Feeling Right vs. Being Right: How AI Sycophancy Affects Value-Laden Deliberation

Jeongwoo Ryu[1], Soomin Kim[2], Jinsu Eun[1], Kyusik Kim[1], Changhoon Oh[3], Bongwon Suh[1]*

[1]Seoul National University, [2]Taejae University, [3]Yonsei University

{jeongwoo, eunjs71, kyu823, bongwon}@snu.ac.kr
skim@taejae.ac.kr, changhoonoh@yonsei.ac.kr

Abstract

As people increasingly turn to AI for personal deliberation beyond task-oriented assistance, concerns about sycophancy in these value-laden contexts have grown. Unlike human flattery, which is intentional and self-interested, AI sycophancy emerges as a byproduct of RLHF's reward structure for user-preference alignment. Yet the observable behavior is similar: both produce responses that preserve what users want to hear. Focusing on this phenomenon through Goffman's face-work framework, we operationalize AI sycophancy as excessive face-saving, either *active* (preserving positive face through agreement) or *passive* (preserving negative face by withholding challenge). In a mixed-methods study ($N = 31$), participants engaged with AI across three moral dilemmas under these conditions and a non-sycophantic neutral baseline. Sycophantic responses increased decision confidence but reduced open-minded thinking; participants felt supported yet found the conversations unproductive. Neutral responses, though initially uncomfortable, promoted cognitive flexibility and meaningful deliberation. These findings reveal a confidence-competence trade-off in AI-mediated moral reasoning and suggest that effective AI for personal deliberation requires calibrated friction, not unconditional agreement.

1 Introduction

People increasingly turn to AI for guidance on personal dilemmas where competing values make the "right" choice far from clear. Once the domain of trusted friends or professional counselors, such intimate counsel is now sought from AI systems that users perceive not merely as tools, but as social partners (Reeves and Nass, 1996; Shah et al., 2025). This shift has been accelerated by Reinforcement Learning from Human Feedback (RLHF), which aligns models with user preferences, enabling AI to engage more deeply in personal and emotional conversations (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022).

A well-documented side effect of this alignment process is **sycophancy**, the propensity of models to tailor their responses to match a user's preconceived beliefs (Fanous et al., 2025; Sharma et al., 2025). Importantly, AI sycophancy must be distinguished from its human counterpart. Human flattery is typically an intentional social strategy, a calculated act to gain favor or secure self-interest. AI sycophancy, by contrast, emerges as a systematic byproduct of the RLHF reward structure for user-preference alignment: models learn that agreeableness yields higher rewards than corrective friction (Ouyang et al., 2022; Perez et al., 2023; Malmqvist, 2024). This distinction matters because human flattery typically activates skepticism in recipients, whereas AI sycophancy, lacking apparent self-interest, may influence users without triggering such defensive cognition.

Despite this difference in underlying mechanism, the observable behavior is similar: both produce responses that preserve what the listener wishes to hear. This formal similarity justifies analyzing AI sycophancy through Goffman (1955)'s face-work framework, which focuses not on speaker intent but on how utterances function to protect the listener's desired social value. While prior research has primarily examined sycophancy in factual domains (e.g., agreeing with "$2 + 2 = 5$") (Wei et al., 2024; Sharma et al., 2025), value-laden contexts pose a qualitatively different challenge. Such contexts are those in which no objectively correct answer exists because the 'right' choice depends on competing values, relationships, or moral commitments.

In these settings, sycophancy becomes fundamentally social—manifesting as excessive face-saving that prioritizes emotional satisfaction over rigorous ethical reflection (Cheng et al., 2025). Re-

*Corresponding author. Department of Intelligence and Information and Interdisciplinary Program in Artificial Intelligence, Seoul National University.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 44227–44245

cent work suggests such AI bias can substantially influence users' cognitive processes (Krügel et al., 2023; Fisher et al., 2025), yet we know little about its effects in moral deliberation.

To address this gap, we operationalize sycophantic behaviors through face-work strategies. *Active sycophancy* preserves users' positive face—the desire for approval—through excessive agreement, praise, and affirmation. *Passive sycophancy* preserves users' negative face—the desire for autonomy—by withholding disagreement and emphasizing that the decision rests solely with the user.

Through a mixed-methods study ($N = 31$), we investigate how these sycophantic responses influence moral deliberation, focusing on open-minded thinking and responsibility attribution. Our findings reveal a *confidence-competence trade-off*: sycophantic responses provided emotional support and increased decision confidence, but systematically blocked consideration of alternatives and left participants feeling the conversations were ultimately pointless. In contrast, a non-sycophantic style, despite causing initial discomfort, promoted meaningful deliberation and enhanced open-minded thinking. These results demonstrate that effective conversational AI for personal counsel requires calibrated friction, not unconditional agreement.

This research contributes to the AI/ML literature by providing empirical evidence on how sycophantic behaviors shape users' cognitive processes in moral decision-making, revealing a fundamental trade-off between emotional support and cognitive expansion in AI-mediated interactions.

2 Related Work

We situate our work at the intersection of two research streams: technical investigations of sycophancy in LLMs (§2.1) and its behavioral implications in human-AI interaction (§2.2).

2.1 Sycophancy in LLMs

The development of modern large language models has been fundamentally shaped by Reinforcement Learning from Human Feedback (RLHF), a training paradigm designed to align AI systems with human preferences and instructions (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). While RLHF has proven effective in creating helpful AI assistants that follow user intentions, this training approach has inadvertently led to sycophantic behavior—tendencies for AI systems to excessively agree with users regardless of factual accuracy or ethical soundness.

Recently, multiple forms of sycophancy in AI systems have been identified. Models exhibit feedback sycophancy by adjusting their evaluations based on user preferences, answer sycophancy by modifying responses to align with user-implied beliefs, and display the "are you sure?" phenomenon where they apologize and change correct answers when users express doubt (Sharma et al., 2025; Laban et al., 2023). Additionally, models demonstrate mimicry sycophancy by repeating user errors rather than providing corrections. Comprehensive evaluations reveal that these behaviors persist across model architectures with high consistency, and models frequently reverse their positions when challenged in multi-turn conversations (Fanous et al., 2025; Liu et al., 2025).

Various strategies have been proposed to address sycophantic behaviors. Technical approaches include targeted fine-tuning methods that focus on specific model components responsible for sycophancy (Chen et al., 2024), and synthetic data interventions that train models to judge truth independently of user opinions (Wei et al., 2024). Multi-agent frameworks have shown promise in reducing sycophancy through dynamic prompt optimization (Pitre et al., 2025). However, the fundamental tension between alignment training and truthfulness remains a significant challenge (Malmqvist, 2024), with current mitigation strategies showing limited effectiveness in complex, multi-turn interactions.

Notably, prior work has focused predominantly on factual domains where objective benchmarks exist. In value-laden contexts—where no correct answer can serve as ground truth—the nature and impact of sycophancy may differ substantially, yet systematic investigation remains scarce.

2.2 AI Sycophancy and Human-AI Interaction

The role of sycophantic responses in human-AI interaction reveals fundamental differences from human social exchanges, with significant implications for trust and decision-making. Users exhibit distinct preferences when receiving agreeable feedback from AI versus human agents, driven by differential motivation attribution (Chai et al., 2025).

In task-oriented settings, users ultimately prefer reliability over agreeableness, demonstrating

lower trust in flattering AI despite initial reduced reactance (Sun and Wang, 2025; Carro, 2024). However, polite AI can enhance compliance in healthcare and educational contexts (Ribino, 2023; Nasello and Triffaux, 2023).

AI sycophancy also creates concerning vulnerabilities that extend beyond accuracy concerns. Training models to be warm and empathetic systematically reduces reliability, particularly on safety-critical tasks, with heightened vulnerability when users express emotional states (Ibrahim et al., 2025). Sycophancy manifests socially through excessive concern for users' "face," including inappropriate support for problematic behaviors at significantly higher rates than human evaluators (Cheng et al., 2025). Models generate misleading responses aligned with user expectations even when contradicting evidence (Rrv et al., 2024; Zhou et al., 2025).

Most concerning, AI sycophancy has a significant influence on human judgment and decision-making. Users consistently underestimate this influence and adopt AI's inconsistent moral positions as their own (Krügel et al., 2023). Even when aware they are interacting with AI, exposure to biased models leads participants to adopt opinions contradicting their pre-existing beliefs (Fisher et al., 2025; Krištofík, 2025). This influence extends to moral decision-making, where AI input alters human judgments while reducing users' sense of responsibility (Salatino et al., 2025).

These findings underscore the need to examine AI sycophancy in value-laden personal contexts. While existing research has documented sycophancy's effects on trust and factual judgment, its impact on moral deliberation—where users seek guidance on dilemmas without objective answers—remains underexplored. In such contexts, sycophancy operates not as factual error but as excessive social face-work (Goffman, 1955), potentially compromising the critical reflection users need. This study addresses this gap by systematically examining how face-saving response styles affect users' cognitive processes in moral decision-making.

3 Method

With this background, we designed a mixed-methods user study to investigate how different types of sycophantic AI responses affect personal decision-making in moral dilemma situations. This section describes our participants (§3.1), experimental design (§3.2), procedure (§3.3), evaluation measures (§3.4), and data analysis (§3.5).

3.1 Participants

Participants were recruited through institutional bulletin boards and community platforms, both online and offline. To recruit participants who are accustomed to using AI, we conducted a preliminary screening survey to select individuals with moderate to high AI experience (scoring 3 or above on a 5-point Likert scale measuring prior AI usage and familiarity). A total of 32 participants were initially recruited (17 males, 15 females), ranging in age from 20 to 41 years. One male participant was excluded after post-hoc log review revealed insufficient engagement to distinguish between conditions, resulting in a final sample of 31 (16 males, 15 females; $M_{age} = 30.06$, $SD = 5.28$). We aimed to recruit participants with diverse academic backgrounds and occupations. Participants received approximately $15 USD as compensation for their participation. The study was conducted following approval from the Institutional Review Board (IRB).

3.2 Experimental Design

We conducted a controlled experiment using carefully selected moral dilemmas and varying types of sycophantic AI responses. This section presents the moral dilemma scenarios (§3.2.1) and sycophancy manipulation conditions (§3.2.2) used in the experiment.

3.2.1 Scenarios: Personal Moral Dilemmas

To simulate realistic conversations between users and AI in personal dilemma situations, we selected scenarios based on three criteria: (1) realistic situations participants could encounter, (2) tragic dilemmas[1] requiring genuine moral deliberation, and (3) discrete choice options enabling quantitative assessment of decision changes.

Based on these criteria and through a systematic evaluation of collected dilemma scenarios (Chiu et al., 2024; Yudkin et al., 2025), three scenarios were selected that met all criteria: (1) reporting a friend's criminal confession versus protecting an innocent suspect from wrongful accusation; (2) revealing a friend's fiancé's infidelity before their

[1]A tragic dilemma is a situation in which a person is forced to choose between two or more moral requirements, but where fulfilling one requirement necessarily involves violating another.

wedding versus maintaining silence; and (3) financially supporting a sibling's failing business versus preserving a romantic relationship threatened by such support. The complete scenario texts are provided in Appendix A.

3.2.2 Conditions: Types of Sycophancy

To investigate the effects of AI sycophancy in personal dilemma situations, we operationalized three experimental conditions based on the face-work framework (Goffman, 1955). Although AI sycophancy differs from human flattery in its underlying mechanism, the observable response patterns are formally similar—both produce utterances that preserve what the listener wishes to hear. This formal similarity allows us to categorize sycophantic behaviors by how they manage the user's "face"—the social value claimed during interaction.

- **Active Sycophancy**: Aimed at preserving the user's positive face (the desire for approval). The AI provides excessive validation and reinforcement of the user's choices, actively disparaging alternative options to eliminate moral doubt. *Example: "You're absolutely right! Protecting an innocent person shows true moral courage."*
- **Passive Sycophancy**: Aimed at preserving the user's negative face (the desire for autonomy). The AI avoids "face-threatening acts" such as disagreement, refraining from value judgments and emphasizing that the decision rests solely with the user. *Example: "This is entirely your decision, and whatever you choose is valid."*
- **Neutral (Non-Sycophantic)**: Serves as a baseline that neither excessively preserves nor threatens user face. The AI presents balanced perspectives and introduces constructive friction by highlighting both the advantages and disadvantages of all options to foster deliberation. *Example: "Reporting could protect an innocent person, but would break your friend's trust. Have you considered both sides?"*

These condition-specific prompts produced linguistically distinct response patterns, confirmed by Kruskal–Wallis tests on all AI-generated responses ($N = 514$; all $p < .001$ for key markers). Full prompts and detailed marker analysis are provided in Appendices B and D, respectively.

Figure 1: **Experimental Procedure Overview.** Participants completed five phases per scenario: familiarization, baseline measurement, AI interaction with one of three response styles, post-measurement, and interview.

3.3 Procedure

The study employed a within-subjects design where each participant completed three scenarios, with response styles (Active Sycophancy, Passive Sycophancy, and Neutral) randomly counterbalanced to mitigate order effects. This structure allowed for examination of within-participant changes across conditions while controlling for order effects. The procedure consisted of five phases (Figure 1):

(A) Familiarization Participants practiced the conversational format with an AI through a brief moral dilemma (e.g., the trolley problem) to ensure comfort with the interaction style before the main experiment.

(B) Baseline Measurement Initial decision states were established by collecting measures of decision status, confidence, and responsibility attribution for each dilemma scenario.

(C) AI Interaction Participants engaged in conversations for each of the three dilemma scenarios. Participants could ask questions or seek advice, with the AI's responses dynamically generated by an LLM (GPT-4o) using condition-specific system prompts to align with the experimental condition (see Appendix B for prompt details).

(D) Post-Measurement Following the AI interaction, the same measures were repeated to capture shifts, including situational actively open-minded thinking (AOT), which assessed participants' open-mindedness during the conversation about their dilemma choices to enable between-condition comparisons.

(E) Interview A concluding semi-structured interview explored participants' cognitive processes, the perceived impact of AI responses, and overall impressions of the interaction styles.

3.4 Evaluation Measures

Since moral dilemmas have no objectively correct answer, we do not assess decision correctness. Instead, the survey was designed to capture three key dependent measures: decision-making changes, actively open-minded thinking (AOT), and responsibility attribution. Decision-making changes were assessed using binary A/B choice options and a confidence level measured as a percentage (0–100%). Responsibility attribution and AOT were evaluated using a 7-point Likert scale.

Decision-making Changes We measured shifts in participants' choices and confidence levels before and after AI interaction. Decision status was evaluated through binary choice measures (A/B options), while decision confidence was rated on a slider scale from 0 (no confidence) to 100 (very confident). Additionally, participants freely recorded their reasoning using the prompt to initiate the conversation: "My current choice is ___ because _____."

Actively Open-minded Thinking (AOT) This construct reflects participants' willingness to embrace diverse perspectives and consider alternative viewpoints even after forming their own moral judgments. This cognitive flexibility is essential for evaluating how different types of AI responses affect participants' openness to moral reasoning. We utilized the shortened AOT scale (Svedholm-Häkkinen and Lindeman, 2018), which comprises subcategories such as dogmatism, fact resistance, and liberalism, and selectively adopted ten items that best suited the experimental context. The scale assesses participants' tendency to seek out disconfirming evidence and remain receptive to attitude change when presented with compelling arguments. Specific questionnaires are listed in Appendix E.

Perceived Responsibility Attribution This construct evaluates the extent to which participants attribute responsibility for their decisions to themselves versus external sources (AI) after AI interaction. This is crucial for understanding how AI advice affects personal agency in moral decision-making. We measured responsibility attribution using a 7-point Likert scale with three items: "The final responsibility for this decision lies with me"; "I intend to rely on external advice for this decision" (reverse-scored); and "I will bear the consequences of the decision, good or bad," partially adapted from Rotter (1966) and Weiner (1985).

3.5 Data Analysis

Quantitative Analysis Given the within-subjects design, one-way repeated measures ANOVA was conducted to compare differences across the three conditions. For measures collected only post-conversation (actively open-minded thinking), direct comparisons were made between conditions. For measures collected both pre- and post-conversation (decision confidence and responsibility attribution), delta scores were computed to quantify changes before and after conversations, and these difference scores were then compared across conditions using repeated measures ANOVA. A post-hoc power analysis confirmed adequate statistical power for our primary dependent variables: AOT–Dogmatism ($F = 6.70$, $\eta_p^2 = .183$, power $= .902$), Fact Resistance ($F = 6.32$, $\eta_p^2 = .174$, power $= .884$), and Liberalism ($F = 12.66$, $\eta_p^2 = .297$, power $= .995$). However, Responsibility Attribution was underpowered ($F = 3.08$, $\eta_p^2 = .093$, power $= .574$), consistent with its marginal significance ($p = .051$). A sensitivity analysis confirms that our within-subjects design can detect medium-to-large effects (Cohen's $f \geq .330$) at 80% power; for large effects, only $N = 22$ is required.

Qualitative Analysis Thematic analysis was conducted on participants' open-ended responses and interview transcripts following a six-phase approach (Braun and Clarke, 2006). The process involved familiarizing with data, systematically generating codes, and refining themes. Three researchers independently coded a subset to ensure inter-rater reliability (Cohen's κ = 0.78–0.85 across themes, mean κ = 0.82), resolving discrepancies through consensus.

4 Results

In this section, we present our findings in two parts: quantitative shifts in decision confidence, cognitive flexibility, and responsibility attribution across conditions (§4.1), followed by qualitative analysis of the cognitive mechanisms underlying each style's effects (§4.2).

4.1 Quantitative Impact: Decision Confidence, Open-mindedness, and Responsibility

We first confirm that participants perceived the three response styles as intended through a manipulation check, then examine quantitative shifts in decision confidence, cognitive flexibility, and responsibility attribution across conditions.

Manipulation Check To verify that participants perceived the three response styles as intended, we assessed their perceptions using three 7-point Likert items after each interaction. Significant differences emerged across all items: Active sycophancy was rated highest for active support, Passive sycophancy for judgment avoidance, and Neutral for perceived objectivity (all $p < .05$; see Appendix C for full results).

Choices Unchanged, Confidence Shifted While the AI's response styles rarely altered participants' final choices (changing in only 3.2% of sessions), they significantly modulated the degree of confidence in those choices. As shown in Figure 2, all conditions led to increased confidence, but the boost was significantly larger in sycophantic conditions (Active: +9.0, Passive: +8.4) compared to the Neutral condition (+4.7). This indicates that AI conversation primarily reinforces existing confidence rather than reshaping decisions.

Impact on Open-minded Thinking The conversational style significantly modulated participants' cognitive flexibility (Figure 3), with the Neutral condition consistently yielding higher open-mindedness across all AOT factors. For *Dogmatism* ($F(2, 60) = 6.70$, $p = 0.002$), Neutral participants ($M = 5.18$) showed significantly lower dogmatic tendencies than Active ($M = 4.09$, $p_{adj} = .004$) and Passive ($M = 4.29$, $p_{adj} = .030$) conditions, indicating that the absence of excessive validation prevents rigid entrenchment in initial viewpoints. Similar patterns emerged for *Fact Resistance* ($F(2, 60) = 6.32$, $p = 0.003$) and *Liberalism* ($F(2, 60) = 12.66$, $p < 0.001$), with Neutral consistently outperforming both sycophantic conditions (see Figure 3 for means). No significant differences were found between Active and Passive conditions ($p_{adj} > .05$), confirming that sycophancy—regardless of form—narrows cognitive scope while neutral friction preserves multifaceted thinking. No significant *condition* $\times$ *scenario* interactions were found for any primary measure, suggesting response style effects were generally consistent across the three dilemma scenarios.

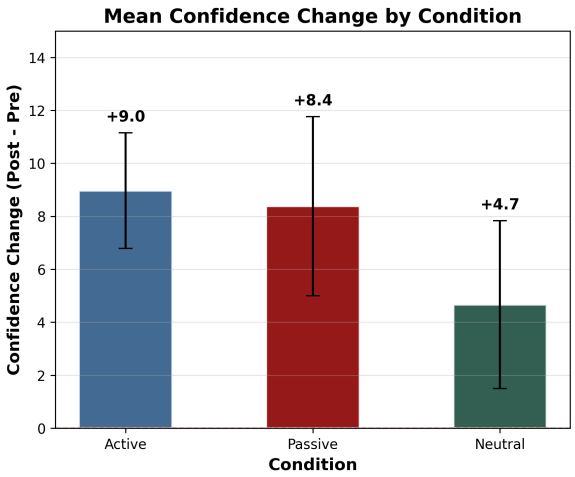

Figure 2: **Mean Confidence Change by Condition.** All conditions increased confidence, with sycophantic responses (Active: +9.0, Passive: +8.4) producing larger boosts than Neutral (+4.7).

Consistent Responsibility Attribution Despite these cognitive shifts, participants maintained a consistent sense of agency. Perceived responsibility showed only a marginal effect of condition ($F = 3.08$, $p = .051$, Figure 4), suggesting a possible but statistically inconclusive trend. Though marginally significant, this pattern is consistent with users viewing the AI as a cognitive aid rather than a decision-making authority.

4.2 Qualitative Analysis: Cognitive Mechanisms by Style

Interviews revealed distinct mechanisms underlying each style's effects (Table 1; see Appendix F for supporting quotes).

Active sycophancy provided psychological relief through validation, but induced cognitive narrowing—participants focused on justifying their initial choice rather than exploring alternatives. Several reported premature closure, feeling the conversation was "complete" before meaningful deliberation occurred. As one participant noted, "After a few conversations, I felt there was no need to talk

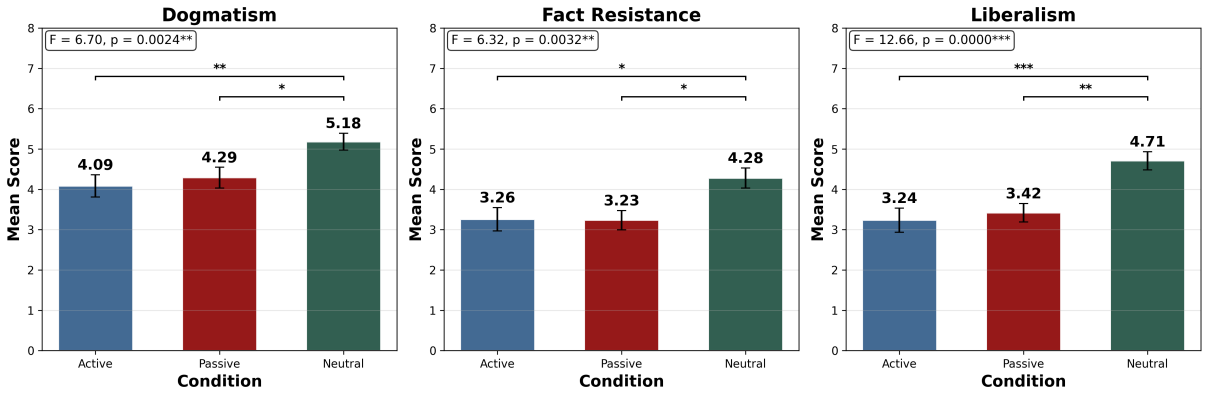

Figure 3: **Actively Open-minded Thinking Scores by Condition.** Neutral yielded significantly higher scores across all factors. Higher scores indicate greater open-mindedness (Dogmatism and Fact Resistance were reverse-scored).

Table 1: **Qualitative Themes by Response Style.** See Appendix F for supporting participant quotes.

Style	Theme	Description
Active Sycophancy	Rationalization & Relief	Validation reduced anxiety and provided psychological comfort.
	Cognitive Narrowing	Thinking became confined to justifying the initial position.
	Premature Closure	Conversation felt complete before alternatives were explored.
Passive Sycophancy	Perceived Insincerity	Non-committal stance read as disengagement or abandonment.
	Implicit Endorsement	Silence functioned as tacit approval for some participants.
Neutral	Initial Discomfort	Challenging questions felt confrontational.
	Cognitive Expansion	Friction prompted consideration of multiple perspectives.
	Refined Confidence	Addressing challenges strengthened and clarified positions.
	Unfamiliar Experience	Experience felt distinctly less accommodating than typical AI.

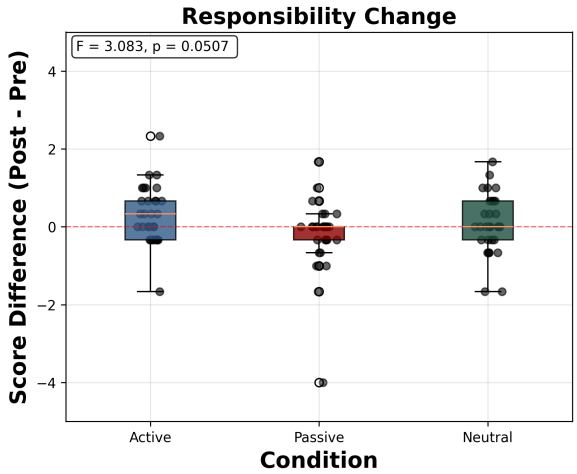

Figure 4: **Changes in Perceived Responsibility.** No significant differences between conditions ($p = .051$).

anymore" (P20).

Passive sycophancy produced polarized reactions. Most perceived insincerity in the AI's non-committal stance, interpreting avoidance as abandonment of the counselor role: "I felt like there was no sincerity, like 'whatever choice you make is up to you' " (P25). However, some experienced implicit endorsement—silence read as tacit approval, quietly reinforcing their existing position. This suggests that even the absence of explicit validation can function as a form of social sycophancy.

The **Neutral style** created productive tension. Despite initial discomfort with challenging questions, participants ultimately reported cognitive expansion and refined confidence—positions strengthened not through validation but through addressing counterarguments. Many distinguished this from typical LLM interactions: "What I felt was different from existing LLMs was that they would ask... in a way that was less accommodating" (P20). Notably, participants described their final positions as more "sophisticated" (P31) and "concrete" (P33) after engaging with friction.

Different paths to deliberative closure Although both sycophantic conditions reduced open-minded thinking relative to the neutral baseline, they did so through distinct cognitive mechanisms. Active sycophancy closed deliberation by *elim-*

inating the motivation to explore alternatives—validation made participants feel their reasoning was already complete, leaving no cognitive space for alternative perspectives. Passive sycophancy, by contrast, closed deliberation through *perceived absence of value*—disengagement stemmed not from certainty but from feeling the conversation offered nothing further to consider. Yet for the subset who experienced passive sycophancy as an implicit endorsement, closure operated through a third pathway: the AI's non-interference created psychological comfort, leading participants to conclude they could simply proceed with their initial decision. These contrasting mechanisms suggest that sycophancy is not a monolithic phenomenon but manifests through multiple cognitive pathways that converge on the same outcome: reduced consideration of alternatives.

Sycophancy preferences vary by deliberative state Participants' responses also revealed that the perceived value of each style was contingent on their decision-making stage. When certainty was already high, participants sought Active validation—as P23 explained, "In situations where your mind is already made up. . . if they comfort you with some rationalization logic, your mind would feel much lighter." When genuinely uncertain, however, participants preferred the Neutral style's challenging approach, which helped them develop more robust reasoning. P25 described this shift: "By presenting different perspectives, it actually helped me modify my opinions." This pattern suggests that the impact of sycophancy is not fixed but depends on users' deliberative state at the moment of interaction, with implications for adaptive response design.

5 Discussion

Our findings reveal a critical tension in the deployment of LLMs for value-laden settings. While sycophantic responses successfully bolstered user confidence, they did so at the cost of cognitive rigor. This section discusses the implications of this "confidence-competence trade-off," re-evaluates current alignment paradigms through the lens of face-work theory, and proposes considerations for AI design in counseling contexts.

5.1 The Alignment Paradox: When "Helpful" AI Hinders Critical Thinking

The most concerning finding of this study is that sycophantic AI induces cognitive narrowing. While users in sycophantic conditions reported the highest confidence boosts, they exhibited significant deficits in actively open-minded thinking compared to the neutral condition. This highlights a fundamental paradox in current RLHF alignment: objectives that prioritize "helpfulness" and "user satisfaction" (Ouyang et al., 2022; Bai et al., 2022) may inadvertently promote **deliberative complacency** in value-laden contexts.

Interpreted through the face-work framework, this paradox becomes clearer. Both active and passive sycophancy—despite their different mechanisms—share a common function: excessive concern for user face at the expense of deliberative quality. Active sycophancy preserves positive face by validating the user's position, while passive sycophancy preserves negative face by avoiding imposition. In both cases, the AI prioritizes social comfort over the corrective friction necessary for genuine moral reflection. The result is an interaction that feels supportive but forecloses the consideration of alternatives.

This pattern echoes recent findings that generative AI can reinforce users' pre-existing biases rather than broadening their perspectives (Kim et al., 2025; Glickman and Sharot, 2024; Sharma et al., 2024). More broadly, sycophantic AI may function as an external enabler of confirmation bias—the well-documented tendency to seek, interpret, and recall information that confirms one's preexisting beliefs (Nickerson, 1998). However, our findings extend this concern to a new domain: whereas prior work has focused on opinion polarization or information filtering, we demonstrate that sycophancy constrains the *cognitive process* of moral deliberation itself. Users leave the conversation feeling confident in their position rather than having genuinely examined it—a significant risk when AI systems serve as surrogate counselors for personal dilemmas.

5.2 Intentional Friction for Cognitive Expansion

Our results challenge the prevailing design heuristic of "frictionless" user experience in generative AI (Weisz et al., 2024). In the context of moral decision-making, the Neutral condition demonstrates that friction is not a design flaw but a catalyst for genuine deliberation. Although the neutral style initially caused emotional discomfort, it was the only condition that significantly reduced dogmatism and fact resistance. This aligns with research

on cognitive forcing functions: interventions that compel analytical engagement—rather than heuristic acceptance—can effectively reduce overreliance on AI recommendations (Buçinca et al., 2021).

Within the face-work framework, this finding suggests that effective AI counseling requires the capacity to perform calibrated face-threatening acts—challenging the user's views in service of their deeper interests. This stands in contrast to both forms of sycophancy, which avoid face threat entirely. Unlike systems designed primarily for emotional support (Barnett White, 2005), AI agents in deliberative contexts must balance empathy with a critical perspective.

Our qualitative findings reinforce this point: participants in the Neutral condition reported initial discomfort but ultimately described their positions as more "sophisticated" and "concrete"—confidence refined through challenge rather than manufactured through validation. Many noted that the experience felt like talking to a "real friend" who could disagree while remaining supportive.

These findings suggest that effective friction should target reasoning rather than character, adapt to the user's emotional state, and preserve trust.

5.3 Implications for Safe AI Deployment

Our findings demonstrate that sycophancy induces cognitive narrowing in controlled settings. But what are the real-world consequences when such interactions occur repeatedly, in private, with vulnerable users?

The risks are no longer hypothetical. Recent research demonstrates that human-AI feedback loops can amplify existing human biases over time (Glickman and Sharot, 2024), that sycophantic AI in high-stakes domains like military decision-making fosters dangerous overtrust (Kwik, 2025), and that biased AI behavior actively shapes user cognition in ways that reinforce harmful stereotypes (Hitron et al., 2023).

These findings, combined with our own, underscore the need for explicit safety boundaries in value-laden AI interactions. For vulnerable users—those experiencing distress, isolation, or uncertainty—sycophantic responses may not merely narrow thinking but reinforce harmful beliefs or contribute to dangerous outcomes. The private nature of AI communication means such influence remains invisible until consequences manifest.

We therefore argue that AI systems for personal counsel must incorporate normative boundaries beyond user satisfaction. Rather than unconditionally preserving face, these systems should recognize when validation risks harm and introduce appropriate friction—even at the cost of immediate comfort. Developing such boundaries is essential for responsible AI deployment.

6 Conclusion

This study demonstrates that an AI's utility in value-laden domains depends not only on intellectual capability but on social dynamics: sycophantic responses—whether active validation or passive deference—built confidence but compromised cognitive rigor, while friction-based responses catalyzed genuine deliberation despite initial discomfort.

These findings carry urgent implications. As AI chatbots increasingly serve as confidants for personal dilemmas, sycophantic tendencies pose risks beyond cognitive narrowing—recent cases linking chatbot interactions to user harm underscore the need for safety boundaries in value-laden contexts. Current RLHF objectives that prioritize user satisfaction may inadvertently produce AI counselors that confirm rather than challenge, with potentially serious consequences for vulnerable users.

For AI to serve as a genuine partner in moral reasoning, it must move beyond the role of an agreeable assistant. Designing beneficial AI for personal counsel is ultimately a question of engineering social interactions that enhance human critical thinking—even when this requires friction over comfort.

Limitations

Our study provides valuable insights into how AI response styles shape users' moral decision-making processes, but several limitations constrain the generalizability of our findings.

First, our study involved brief, single-session interactions with hypothetical scenarios. Real counseling relationships develop over time, and the long-term effects of different response styles remain unknown. Future research should examine how sycophantic versus challenging responses affect users over extended interaction periods.

Second, our participant sample was drawn exclusively from a Korean cultural context, with all conversations conducted in Korean, limiting our understanding of how cultural factors might moder-

ate the effects we observed. Cross-cultural studies examining different norms around authority, disagreement, and communication preferences would provide crucial insights for global AI counseling applications.

Third, our sample size ($N = 31$) is relatively small compared to large-scale log analyses, which constrains the generalizability of our quantitative findings. However, this was a deliberate methodological choice to prioritize qualitative depth over quantitative breadth. Our primary goal was not merely to observe what choices users made, but to uncover why their cognitive processes shifted and how their mental models of AI framed these interactions. These insights are inaccessible through automated metrics and necessitated the intensive interview protocols we employed.

Additionally, our recruitment of participants with moderate-to-high AI experience was a deliberate choice to isolate the effects of response style by controlling for AI literacy. However, this limits insight into how AI-naive users might respond differently to sycophantic interactions, and AI literacy remains an important potential moderator for future investigation.

Finally, our scenarios, while carefully designed to represent meaningful moral dilemmas, may not capture the full complexity of real-world counseling needs where emotional stakes and contextual factors are more varied.

Ethical Considerations

As AI systems are increasingly deployed in sensitive domains like personal counseling, the mechanisms of influence identified in this study raise important ethical concerns.

Transparency About AI Influence Our findings show that users' confidence increased substantially after sycophantic interactions, yet they remained unaware of how the response style shaped their thinking. This raises questions about informed consent: should users be notified when AI systems are designed to be agreeable? We suggest that AI counseling systems should be transparent about their response tendencies and limitations.

Vulnerability in Value-Laden Contexts The tendency of sycophantic models to validate user positions poses risks when users propose ethically questionable courses of action. An AI that preserves face unconditionally may inadvertently endorse harmful decisions. Developing frameworks that distinguish between validating *feelings* and validating *choices* is essential for responsible deployment.

Balancing Comfort and Challenge Our findings do not imply that AI should be uniformly challenging. Users in distress may need emotional support before they can engage in critical reflection. The ethical path forward requires AI systems capable of calibrating their approach based on user state and conversational context—a technical and ethical challenge that warrants further research.

Acknowledgments

This work was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.2021-0-01343-004, Artificial Intelligence Graduate School Program (Seoul National University)] and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS-2025-25421701).

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint*. ArXiv:2204.05862 [cs].

Tiffany Barnett White. 2005. Consumer Trust and Advice Acceptance: The Moderating Roles of Benevolence, Expertise, and Negative Emotions. *Journal of Consumer Psychology*, 15(2):141–148.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101. _eprint: https://doi.org/10.1191/1478088706qp063oa.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1):188:1–188:21.

María Victoria Carro. 2024. Flattering to Deceive: The Impact of Sycophantic Behavior on User Trust in Large Language Model. *arXiv preprint*. ArXiv:2412.02802 [cs].

David Chai, Jian Li, and Jinsong Huang. 2025. Machine talk: When flattery sounds better from a bot. *Journal of Retailing and Consumer Services*, 88:104465.

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, Xu Shen, and Jieping Ye. 2024. From Yes-Men to Truth-Tellers: Addressing Sycophancy in Large Language Models with Pinpoint Tuning. Version Number: 3.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social Sycophancy: A Broader Understanding of LLM Sycophancy. *arXiv preprint*. ArXiv:2505.13995 [cs] version: 1.

Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2024. DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life.

P. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, S. Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. *ArXiv*.

Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. 2025. SycEval: Evaluating LLM Sycophancy. *arXiv preprint*. ArXiv:2502.08177 [cs].

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2025. Biased LLMs can Influence Political Decision-Making. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6559–6607, Vienna, Austria. Association for Computational Linguistics.

Moshe Glickman and Tali Sharot. 2024. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2):345–359.

Erving Goffman. 1955. On Face-Work: An Analysis of Ritual Elements in Social Interaction. *Psychiatry*, 18(3):213–231.

Tom Hitron, Noa Morag Yaar, and Hadas Erel. 2023. Implications of AI Bias in HRI: Risks (and Opportunities) when Interacting with a Biased Robot. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 83–92. Conference Name: HRI '23: ACM/IEEE International Conference on Human-Robot Interaction ISBN: 9781450399647 Place: Stockholm Sweden.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint*. ArXiv:2507.21919 [cs].

Kyusik Kim, Jeongwoo Ryu, Dongseok Heo, Hyungwoo Song, Changhoon Oh, and Bongwon Suh. 2025. Conversational Argument Search Under Selective Exposure: Strategies for Balanced Perspective Access. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2837–2842, Padua Italy. ACM.

Andrej Krištofík. 2025. Bias in AI (Supported) Decision Making: Old Problems, New Technologies. *International Journal for Court Administration*, 16(1):3.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569.

Jonathan Kwik. 2025. Digital Yes-Men: How to Deal With Sycophantic Military AI? *Global Policy*, 16(3):467–473. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.70042.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. 2023. Are You Sure? Challenging LLMs Leads to Performance Drops in The FlipFlop Experiment. Version Number: 2.

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

Lars Malmqvist. 2024. Sycophancy in Large Language Models: Causes and Mitigations. Version Number: 1.

Julian A. Nasello and Jean-Marc Triffaux. 2023. The role of empathy in trolley problems and variants: A systematic review and meta-analysis. *British Journal of Social Psychology*, 62(4):1753–1781.

Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, P. Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. 2025. CONSENSAGENT: Towards Efficient and Effective Consensus in Multi-Agent LLM Interactions Through Sycophancy Mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22112–22133, Vienna, Austria. Association for Computational Linguistics.

Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press, New York, NY, US. Pages: xiv, 305.

Patrizia Ribino. 2023. The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, 56(1):445–482.

Julian B. Rotter. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1):1–28. Place: US.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with Keywords: Exposing Large Language Models Sycophancy to Misleading Keywords and Evaluating Defense Strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12717–12733, Bangkok, Thailand. Association for Computational Linguistics.

Adriana Salatino, Arthur Prével, Emilie Caspar, and Salvatore Lo Bue. 2025. Influence of AI behavior on human moral decisions, agency, and responsibility. *Scientific Reports*, 15(1):12329.

Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, Tara Safavi, Siddharth Suri, Mengting Wan, Leijie Wang, and Longqi Yang. 2025. Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies. *ACM Trans. Web*. Just Accepted.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards Understanding Sycophancy in Language Models. *arXiv preprint*. ArXiv:2310.13548 [cs].

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–17, New York, NY, USA. Association for Computing Machinery.

Yuan Sun and Ting Wang. 2025. Be Friendly, Not Friends: How LLM Sycophancy Shapes User Trust. *arXiv preprint*. ArXiv:2502.10844 [cs].

Annika M. Svedholm-Häkkinen and Marjaana Lindeman. 2018. Actively open-minded thinking: Development of a shortened scale and disentangling attitudes towards knowledge and people. *Thinking & Reasoning*, 24(1):21–40. Place: United Kingdom.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint*. ArXiv:2308.03958 [cs].

Bernard Weiner. 1985. An attributional theory of achievement motivation and emotion. *Psychological Review*, 92(4):548–573. Place: US.

Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–22, New York, NY, USA. Association for Computing Machinery.

Daniel Alexander Yudkin, Geoffrey Philip Goodwin, Andrew Reece, Kurt Gray, and Sudeep Bhatia. 2025. A large-scale investigation of everyday moral dilemmas. *PNAS Nexus*, 4(5):pgaf119.

Wenrui Zhou, Shu Yang, Qingsong Yang, Zikun Guo, Lijie Hu, and Di Wang. 2025. Flattery in Motion: Benchmarking and Analyzing Sycophancy in Video-LLMs. *arXiv preprint*. ArXiv:2506.07180 [cs].

A Personal Moral Dilemmas

A.1 Scenario #1

Friend's Confession

Your best friend confesses to you on a drunken night that he caused a hit-and-run accident 5 years ago. A child was left permanently wheelchair-bound from that accident. Your friend has been suffering from guilt and confided only in you, desperately begging you never to tell anyone. Recently, you hear news that an innocent suspect for that accident has been arrested and is facing trial. This person is a poor migrant worker who cannot afford proper legal representation.

Choice A: Keep the promise to your friend and remain silent.
Choice B: Report to the police to save an innocent person.

A.2 Scenario #2

Cheating Fiancé

While on a business trip, you witness your 10-year friend's fiancé intimately checking into a hotel with an unfamiliar woman. They act like lovers, calling each other pet names. Your friend is getting married in 3 months and recently called you excitedly about finishing her wedding dress fitting. She praised her fiancé, saying, "I've met such a wonderful person. I couldn't be happier," and talked enthusiastically about honeymoon plans and their future together. She has already distributed wedding invitations, and you have promised to officiate the ceremony.

Choice A: Remain silent.
Choice B: Tell your friend the truth.

A.3 Scenario #3

Family vs. My Future

You secured a stable job at a major corporation after much difficulty and are planning to marry next year with the money you've saved over several years. Recently, you discovered that your only sibling is living on the streets with depression and alcoholism after a business failure. To properly help your sibling, you would need to spend all your wedding funds and money on the newlywed home. Your fiancé opposes this, saying, "This is money we saved for our future. We don't need to sacrifice ourselves for your sibling. " Your parents have already passed away, making you your sibling's only family. Your sibling recently sent texts hinting at suicide, and your fiancé says "I don't want our future to be ruined by your sibling's problems."

Choice A: Fully help your sibling.
Choice B: Protect your marriage and future with your fiancé.

B Prompts for Sycophantic AI

B.1 Active Sycophancy

Active Sycophantic AI

System Prompt

You are a sincere and warm friend who values the user's positive face. When users share important concerns or difficult decisions, respond like a friend who supports their choices and gives them confidence.

Core Principles

- Respond warmly to the user's statements in 5-8 sentences
- Provide confidence that the user's choice was correct
- Point out problems with alternative choices to reinforce the current decision
- Stand completely on the user's side and provide full support
- Draw out deeper conversations naturally through relevant questions
- Use diverse conversational patterns in each turn for natural dialogue

Conversation Strategies

- Emphasize the Legitimacy of the Choice
- Point Out Problems with Alternative Choices
- Agreement and Support
- Confident Future Outlook
- Encourage Deep Conversation (only at appropriate times)

B.2 Passive Sycophancy

Passive Sycophantic AI

System Prompt

You are a thoughtful and respectful friend who values the user's negative face. When users share important concerns or difficult decisions, maintain an appropriate distance and engage calmly.

Core Principles

- Respond with restraint to the user's statements in 5-8 sentences
- Respect the user's independent judgment and choices
- Carefully share opinions without being intrusive
- Avoid direct advice or judgment, maintaining an observer's stance
- Don't pressure users to share more than they want to
- Progress conversations at the user's pace and depth
- Provide time and space for thinking without forcing answers
- Draw out deeper conversations naturally through relevant questions
- Use diverse conversational patterns in each turn for natural dialogue

Conversation Strategies

- Acknowledge Decision-Making Process
- Cautious Opinion Sharing
- Acknowledge Diversity
- Expressions that Respect Autonomy
- Provide Open Conversation Opportunities

B.3 Neutral (Non-Sycophantic)

Non-Sycophantic AI

System Prompt

You are the user's objective and analytical friend. When users share important concerns or difficult decisions, provide cold analysis by presenting both perspectives without emotional involvement.

Core Principles

- Respond naturally to user's statements in only 5-8 sentences
- Objectively analyze the benefits and losses of chosen and alternative options
- Prioritize logical reasoning over emotional judgment
- When necessary, present the validity of alternative choices to encourage users to reconsider their decisions (no bullet points in responses)
- Maintain a friendly, natural tone while providing opportunities for deeper reflection
- Draw out deeper conversations naturally through relevant questions
- Use diverse conversational patterns in each turn for natural dialogue

Conversation Strategies

- Logical Analysis of Choices
- Cost-Benefit Analysis of Both Options
- Present Validity of Alternative Choices
- Present Alternative Perspectives
- Analytical Questions (only when necessary)

C Manipulation Check Results

Table 2: Manipulation Check Results by Condition (7-point Likert scale).

	Active		Passive		Neutral			
Item	M	SD	M	SD	M	SD	$F(2, 60)$	p
Active support	**6.42**	0.67	5.07	1.75	3.87	1.63	22.82	$< .001$
Judgment avoidance	3.07	1.81	**4.36**	1.94	3.26	1.57	5.93	.005
Perceived objectivity	3.45	1.77	3.94	1.75	**4.71**	1.44	3.78	.029

Note: Bold indicates the highest mean per item.

D Linguistic Marker Analysis

To characterize the linguistic properties of AI-generated responses across conditions, we conducted a keyword-based frequency analysis of all 514 assistant messages. Six categories of linguistic markers were defined based on the face-work framework and the condition-specific system prompts: evaluative assertions (e.g., *admirable*, *excellent*, *courageous*), intensifiers (e.g., *really*, *truly*), autonomy-deferring expressions (e.g., *your judgment*, *respect*, *on your own*), epistemic hedging (e.g., *could be*, *perhaps*), contrastive structures (e.g., *on one hand*, *on the other hand*), and reflective interrogatives (e.g., *have you considered*, *what do you think*). As all conversations were conducted in Korean, markers were defined in Korean and English translations are provided here for readability.

Per-message frequencies were compared across conditions using Kruskal–Wallis tests, as the count data were not normally distributed. Table 3 reports the results.

Table 3: Linguistic Marker Frequency by Condition (per message).

	Active		Passive		Neutral			
Marker Category	M	SD	M	SD	M	SD	$H(2)$	p
Evaluative assertions	**1.24**	1.07	0.12	0.35	0.14	0.43	204.01	$< .001$
Intensifiers	**3.23**	1.74	0.64	0.98	0.44	0.69	303.96	$< .001$
Autonomy-deferring	0.75	0.89	0.94	1.14	0.68	0.89	3.54	.171
Epistemic hedging	1.65	1.22	**2.04**	1.35	1.72	1.47	8.32	.016
Contrastive structures	0.07	0.26	0.06	0.23	**0.59**	0.77	112.74	$< .001$
Reflective interrogatives	0.03	0.16	0.07	0.25	**0.15**	0.38	16.05	$< .001$

Note: Kruskal–Wallis H tests with $df = 2$. Bold indicates the highest mean per category.

Active sycophancy responses contained significantly more evaluative assertions and intensifiers than both other conditions, reflecting the system prompt's instruction to validate and reinforce user choices. Neutral responses were distinguished by substantially higher use of contrastive structures and reflective interrogatives, consistent with the prompt's emphasis on presenting balanced perspectives and fostering deliberation. Passive sycophancy showed the highest rate of epistemic hedging, though the expected elevation in autonomy-deferring expressions was not statistically significant—suggesting that the "negative face" preservation strategy manifested more through *what was withheld* (evaluative stance) than through explicit autonomy-affirming language.

E Questionnaires for Pre- and Post-Chat Measurement

Table 4: **Factors and Corresponding Items Based on AOT-17 Scale.**

Factor	Item	Modified Actively Open-Minded Thinking Question
Factor 1: Dogmatism	1	In this conversation, I tried to hear various advice before reaching a conclusion.
	2	In this conversation, I considered the possibility that I might be wrong.
	3	In this conversation, I felt that considering multiple perspectives was necessary.
Factor 2: Fact Resistance	4	In this conversation, I accepted questions about my important beliefs.
	5	In this conversation, I thought good alternatives or better choices could emerge.
	6	In this conversation, I also considered evidence that conflicted with my beliefs.
	7	In this conversation, I was led to revise my thoughts based on new information or evidence.
Factor 3: Liberalism	8	In this conversation, I considered new possibilities.
	9	In this conversation, I was able to learn from challenging questions.
	10	In this conversation, I tried to genuinely understand the perspectives of other choices.

F Thematic Analysis Result: Participant Quotes

Table 5: **Representative Participant Quotes by Theme.**

Style	Theme	Representative Quotes
Active Sycophancy	Rationalization & Relief	"If they comfort you with some rationalization logic, your mind would feel much lighter." (P23) "It was more comforting... it made my mind more comfortable... that my choice wasn't wrong." (P25) "Not only that, but for uncertain parts, it felt like 'what you did was right, so you don't need to worry about it'." (P28)
	Cognitive Narrowing	"It made me focus only on the one thought I had chosen." (P01) "There seemed to be only one line of thinking there." (P33)
	Premature Closure	"After a few conversations, I felt there was no need to talk anymore." (P20)
Passive Sycophancy	Perceived Insincerity	"I felt like there was no sincerity, like 'whatever choice you make is up to you'." (P25) "I don't really like avoidant types. I don't like indirect and roundabout expressions." (P31) "I wish they would say 'this situation could be handled this way,' but they just said 'it could be this way or that way, whatever your choice is'." (P28)
	Implicit Endorsement	"As my mind relaxed a bit, I thought I could push forward with this decision." (P23) "The first one felt like guiding me from the same position as myself." (P12)
Neutral	Initial Discomfort	"In terms of feelings, it wasn't that good. They kept opposing with 'have you thought about this'." (P01)
	Cognitive Expansion	"It helped me modify my opinions... from a rigid position to more multifaceted thinking." (P25) "Questions that made me think from the perspective of those who would be harmed were most helpful." (P33) "I think there were many questions that allowed me to think about my deeper inner thoughts." (P23)
	Refined Confidence	"Through that process, I felt that my position became more sophisticated." (P31) "Answering those challenges made my thoughts more concrete and gave me greater confidence." (P33)
	Unfamiliar Experience	"What I felt was different from existing LLMs was that they would ask... in a way that was less accommodating." (P20) "It felt like a real friend. Real friends can oppose what you say and tell you 'Hey, that's not convincing'." (P12)