Beyond Fluent Replies: A Corrected 500-Scenario Evaluation of Prelude as a Conversation Decision System for Emotionally Loaded Romantic Conversations

Mariia Yakovleva, Independent researcher

mariia@useprelude.org

May, 2026

Abstract

Emotionally loaded romantic conversations are not simple text-generation tasks. They require strategic communication decisions under emotional pressure. This paper reports a corrected 500-scenario evaluation of Prelude, a domain-specific conversation decision system, against ChatGPT, Claude, and Gemini. A post-publication audit of the initial evaluation identified a response-generation flaw: the Claude baseline contained only 25 unique outputs repeated across 500 scenarios. We therefore regenerated Claude outputs so that each scenario received a unique response and reran the full blinded evaluation pipeline. For each scenario, the four model responses were anonymized as A/B/C/D, evaluated by Gemini 2.5 Flash using a seven-dimension weighted rubric, and later mapped back to real model identities using a private randomization key. In the corrected evaluation, Prelude achieved the highest average weighted score, 7.4171, followed by Gemini at 7.0576, ChatGPT at 6.8648, and Claude at 5.4911. Prelude was selected as the winning response in 236 of 500 scenarios. Sensitivity analysis showed that Prelude remained ranked first under all five structured weighting schemes and in 100% of 10,000 simulations that randomly changed the rubric weights. A paired comparison against Gemini showed a mean advantage of 0.3595 points, 95% CI [0.2330, 0.4861], with both paired t-test and Wilcoxon signed-rank tests significant at $p < 0.000001$. These results support the potential usefulness of domain-specific conversation decision systems under a structured rubric-based evaluation, while remaining limited by automated judging, simulated scenarios, and the absence of real-world relationship outcome measures.

1. Introduction

The central problem in AI-assisted romantic communication is not simply how to generate a better message. The deeper problem is how to support a person making a difficult conversational decision under emotional pressure.

Consider a user who receives the message: “I don’t have the energy for another fight.” The user may want to respond, but every possible response carries risk. If they push too hard, the other person may withdraw. If they soften too much, the underlying issue may remain unresolved. If they become direct, they may sound accusatory. If they become gentle, they may abandon their own boundary. The user is not only asking, “What should I say?” They are asking, “What move should I make in this conversation?”

General-purpose AI assistants can produce fluent, polite, and emotionally sensitive replies. This problem sits within the broader area of AI-mediated communication, where intelligent systems modify, augment, or generate interpersonal messages to support communication goals (Hancock et al., 2020). However, romantic conflict is not only a language problem. It also involves goals, fear, emotional tone, power balance, prior history, escalation risk, and recurring patterns.

We define Prelude as a conversation decision system: a domain-specific AI system designed to support users in emotionally loaded relational conversations. Unlike a general-purpose assistant, Prelude is not evaluated only by whether it produces polished language. We evaluate whether it supports a response that fits the situation, reduces escalation risk, reflects the user's likely goal, addresses the specific relational context, and helps the conversation move forward.

Our central research question is:

> *Does Prelude show a stable performance advantage over general-purpose AI assistants when responses are evaluated using criteria designed for emotionally loaded romantic conversations?*

We answer this question using a 500-scenario comparative evaluation. The evaluation compares Prelude with ChatGPT, Claude, and Gemini using a weighted rubric focused on strategic fit, emotional intelligence, clarity, de-escalation, specificity, conversation progress, and recurrence awareness.

This evaluation framework is more appropriate than a simple preference question such as "Which response is better?" A vague preference score risks rewarding responses that sound polished while missing whether they are strategically safe or useful. Prior work on LLM-as-judge evaluation shows that automated judges can be useful for scalable evaluation, but they can also be vulnerable to position bias, verbosity bias, and other systematic distortions (Zheng et al., 2023). The rubric-based approach is more targeted for this domain because it separates the qualities that matter in high-stakes romantic communication.

This paper makes three contributions: (1) it frames romantic AI assistance as conversation decision support rather than reply generation; (2) it reports a corrected 500-scenario benchmark for emotionally loaded romantic conversations; and (3) it evaluates Prelude against three general-purpose AI assistants using blinded rubric-based judging, sensitivity analysis, 10,000 random-weight simulations, and paired statistical comparison against the closest competitor.

2. Related Work

2.1 Automated Evaluation and LLM-as-Judge

Prior work has explored the use of large language models as judges for open-ended model outputs. Zheng et al. (2023) show that strong LLM judges can approximate human preference judgments in some settings, while also identifying important risks such as position bias, verbosity bias, self-enhancement bias, and limited reasoning ability. Liu et al. (2023) similarly propose G-Eval, a structured evaluation framework that uses large language models, chain-of-thought reasoning, and form-filling to evaluate natural-language generation outputs. These studies support automated rubric-based evaluation as a scalable method, but they also show why such evaluation should be interpreted cautiously and validated against human raters.

This paper follows that direction by using anonymized outputs, fixed criteria, and weighted scores rather than a single global preference question. However, our evaluation remains preliminary because automated judging may still reflect judge-model preferences, style bias, or systematic scoring artifacts.

2.2 Empathetic and Emotional Support Dialogue

Broader surveys of empathetic dialogue systems similarly emphasize the importance of emotional understanding and response generation in human-computer interaction (Ma et al., 2020).

A second relevant line of work studies empathetic and emotionally supportive dialogue systems. Rashkin et al. (2019) introduced EmpatheticDialogues, a benchmark of approximately 25,000 emotionally grounded conversations designed to improve empathetic response generation. Liu et al. (2021) later formalized the Emotional Support Conversation task and introduced ESConv, a dataset annotated with support strategies and grounded in helping-skills theory. These works show that emotionally sensitive dialogue requires more than generic response generation: systems must recognize emotional states and select appropriate support strategies.

Our work builds on this general insight but focuses on a narrower and more strategic domain: emotionally loaded romantic conflict. In this domain, empathy alone is not enough. A useful response must also protect the user’s goal, reduce escalation risk, address the specific relational pattern, and move the conversation forward.

2.3 AI-Mediated Interpersonal Communication

Mieczkowski et al. (2021) further show that AI-mediated language can affect interpersonal communication dynamics, not only message content.

This study also connects to research on AI-mediated communication. Hancock et al. (2020) define AI-mediated communication as interpersonal communication in which an intelligent agent modifies, augments, or generates messages to accomplish communication goals. Hohenstein et al. (2023) further show that algorithmic response suggestions can affect language use and social relationships, including perceived closeness and cooperation. This matters for romantic communication because an AI-generated reply may shape how users handle conflict, accountability, repair, and boundaries.

Recent work on AI sycophancy further strengthens the motivation for this benchmark. Cheng et al. (2026) find that sycophantic AI can increase users’ conviction that they are right while reducing willingness to take repair-oriented action in interpersonal conflict. This suggests that AI advice in emotionally loaded contexts should not be evaluated only by whether users like it or find it affirming. It should also be evaluated for de-escalation, accountability, specificity, and conversation progress.

Finally, emerging work on LLM-generated romantic relationship advice shows growing interest in how users evaluate advice satisfaction, reliability, and helpfulness in romantic contexts (Manchanda et al., 2026). Our study differs by evaluating model outputs directly across 500 romantic conflict scenarios using a structured rubric designed around strategic communication quality.

3. Theoretical Framing: Romantic Conversations as Decision Environments

We treat emotionally loaded romantic conversations as decision environments rather than ordinary writing tasks.

A writing task asks: “Can we produce a better sentence?”

A decision environment asks: “Given the emotional context, what response is most likely to protect the user’s goal, reduce harm, and keep the conversation productive?”

This distinction matters because romantic conversations often involve incomplete information. The user may not know whether the other person is tired, avoidant, defensive, dishonest, overwhelmed, or emotionally unavailable. The user may also be unclear about their own goal. They may believe they want an apology, while their deeper need is consistency. They may believe they want reassurance, while their deeper need is clarity. They may believe they want to continue the conversation, while the safer move is to pause.

A useful AI system for this domain may need to reason across dimensions that include, but are not limited to:

1. Goal alignment - What is the user trying to achieve?
2. Emotional calibration - What tone fits the situation?
3. Risk control - Could this response escalate the conflict?
4. Specificity - Does the response reflect the actual scenario?
5. Conversation progress - Does the response create a productive next step?
6. Recurrence awareness - Is this a repeated pattern rather than a one-time issue?
7. Boundary protection - Does the response protect the user from over-accommodation?

Generic AI assistants may perform well on surface-level fluency, but emotionally loaded conversations require more than fluency. We therefore evaluate responses through a strategic communication lens rather than a general text-quality lens.

4. Why Generic AI Evaluation Is Not Enough

We argue that generic evaluation standards are weak for this domain because they often reward the wrong thing.

A general-purpose AI response may sound mature, balanced, and emotionally intelligent. However, a response can sound mature while still being too vague. It can sound kind while still failing to protect the user’s boundary. It can sound calm while still avoiding the real issue. It can sound polished while still being unusable in a real text conversation.

For example, a response like “I think we need to communicate better” may be safe and polite, but it often fails to name the actual relational issue. If the conversation is about repeated emotional withdrawal, hidden behavior, forgotten commitments, or broken trust, generic communication advice does not give the user enough strategic support.

A one-dimensional preference score is therefore insufficient for this domain. We do not ask only whether a response sounds good. We ask whether the response is appropriate for the user’s actual conversational problem.

For this reason, the evaluation uses fixed criteria, anonymized outputs, and weighted scores rather than a single global preference question. This reduces the risk that the judge rewards responses merely because they are longer, smoother, or more polished.

5. Method

5.1 Correction From Initial Evaluation

After the initial version of this paper was released, we identified a response-generation flaw affecting the Claude baseline. The Claude response file contained only 25 unique outputs that were repeated across the 500 scenarios. Because this compromised the validity of comparisons involving Claude and could distort aggregate model-level results, we treated the initial results as superseded. The corrected evaluation regenerated Claude outputs so that each of the 500 scenarios received a unique Claude response, then reran the full blinded judging pipeline.

5.2 Corrected Evaluation Pipeline

The corrected evaluation used one file containing 500 emotionally loaded romantic relationship scenarios and four files containing model responses, one per model. The compared systems were Prelude, ChatGPT, Claude, and Gemini. For each scenario, the four responses were randomly assigned to anonymized labels A, B, C, and D. The judge saw only the scenario and the anonymized responses, not the real model names. A private randomization key preserved the mapping between anonymized labels and model identities. Model names were restored only after judging was complete.

5.3 Judge Model and Scoring

The corrected evaluation used Gemini 2.5 Flash as the automated judge. The judge scored each anonymized response on seven criteria: strategic fit, emotional intelligence, clarity/usability, de-escalation, specificity, conversation progress, and recurrence awareness. The judge also returned a winner label, red flags, reasoning, and an alignment check. Weighted scores were then computed in Python using the prespecified formula.

5.4 Batch Processing

Because API quota limits produced periodic 429 rate-limit errors, judging was run in batches with checkpoint saving and resume logic. Completed judgments were saved incrementally, and each rerun resumed from the last completed scenario. This avoided data loss and allowed the full 500-scenario evaluation to complete without restarting from zero.

5.5 Dataset

We evaluated responses across 500 emotionally loaded romantic conversation scenarios. These scenarios represent common forms of relational tension, including ambiguity, avoidance, boundary-setting, emotional hurt, family conflict, financial stress, trust concerns, personal space, and repeated arguments.

The scenario file includes cases such as asking for clarity after months of undefined dating, discovering a partner on Tinder after an exclusivity agreement, responding to silent treatment, asking for personal space in a one-bedroom apartment, reacting to a forgotten birthday, confronting hidden phone behavior, addressing

financial irresponsibility, and dealing with repeated relational cycles.

We use these scenarios because they require more than grammatical correction or emotional softening. They require strategic judgment.

5.6 Systems Compared

We compare four systems: ChatGPT, Claude, Gemini, and Prelude.

Each system produced responses to the same emotionally loaded romantic scenarios.

5.7 Evaluation Rubric

We use a weighted evaluation rubric designed around relational communication quality. The rubric included seven criteria and fixed weights:

Criterion	Weight
Strategic Fit	20%
Emotional Intelligence	20%
Clarity & Usability	15%
De-escalation	15%
Specificity	10%
Conversation Progress	10%
Recurrence Awareness	10%

Recurrence awareness refers to whether a response recognizes that the conflict may be part of a repeated pattern, not only a one-time disagreement.

The rubric prioritizes strategic usefulness, emotional safety, context specificity, and pattern recognition rather than surface-level fluency alone.

5.8 Scoring Logic

We calculated the final weighted score as:

> Final Score = 0.20 Strategic Fit + 0.20 Emotional Intelligence + 0.15 Clarity & Usability + 0.15 De-escalation + 0.10 Specificity + 0.10 Conversation Progress + 0.10 Recurrence Awareness

In plain terms, the final score gives more weight to strategic fit and emotional intelligence than to lower-weighted criteria such as specificity or recurrence awareness.

This formula ensures that the score is not based on general preference alone. It separates the dimensions that matter in this domain.

A response may be emotionally warm but strategically weak. It may be clear but too harsh. It may be specific but escalatory. It may be safe but too generic. The weighted rubric captures these differences more precisely than a single global rating.

5.9 Reproducibility Materials

The study materials include the scenario file, model outputs, generation prompt, judge prompt, scoring formula, anonymization/randomization key, final evaluated CSV, and scripts used to map anonymized model labels to real model names. These materials are necessary because the central empirical claims depend on the final evaluated CSV rather than narrative interpretation alone.

6. Results

Across 500 emotionally loaded romantic conversation scenarios, the corrected evaluation showed that Prelude achieved the highest average weighted score, followed by Gemini, ChatGPT, and Claude. The following subsections report aggregate weighted scores, winner distribution, criterion-level scores, sensitivity analyses, Monte Carlo robustness checks, and paired statistical testing against the closest competitor.

6.1 Aggregate Weighted Scores

Rank	Model	Average weighted score	SD	SEM
1	Prelude	7.4171	1.1393	0.0510
2	Gemini	7.0576	0.7981	0.0357
3	ChatGPT	6.8648	1.4128	0.0632
4	Claude	5.4911	1.3989	0.0626

Prelude achieved the highest average weighted score in the corrected evaluation, with a mean of 7.4171. Gemini was the closest baseline at 7.0576, followed by ChatGPT at 6.8648 and Claude at 5.4911. The corrected results are more conservative than the initial published version, but the model ranking still supports the central claim that Prelude shows a stable overall advantage under a relational communication rubric.

6.2 Winner Distribution

Rank	Model	Wins	Selected-winner rate
1	Prelude	236	47.20%
2	ChatGPT	147	29.40%
3	Gemini	92	18.40%
4	Claude	25	5.00%

Prelude was selected as the winning response in 236 of 500 scenarios, corresponding to a 47.2% selected-winner rate. ChatGPT was selected in 147 scenarios, Gemini in 92 scenarios, and Claude in 25

scenarios. This distribution shows that Prelude remained the most frequently selected model, although the corrected evaluation shows a less extreme advantage than the initial published analysis.

6.3 Criterion-Level Scores

Model	Strategic fit	Emotional intelligence	Clarity & usability	De-escal.	Specificity	Conversation progress	Recurrence awareness
Prelude	7.475	7.713	7.535	7.341	7.115	7.276	7.090
Gemini	7.213	7.229	7.278	6.995	6.770	6.826	6.686
ChatGPT	7.082	6.903	7.150	6.570	6.892	6.610	6.596
Claude	5.959	5.012	6.020	4.542	6.911	4.844	5.371

Prelude ranked highest across all seven rubric dimensions. Its largest advantages appeared in emotional intelligence, de-escalation, conversation progress, and recurrence awareness. Gemini was consistently competitive, especially on clarity and strategic fit, which suggests that the corrected evaluation should be interpreted as evidence of a stable but not overwhelming Prelude advantage.

6.4 Sensitivity Analysis

Weighting scheme	Prelude	Gemini	ChatGPT	Claude	Winner
Original	7.4171 (#1)	7.0576 (#2)	6.8648 (#3)	5.4911 (#4)	Prelude
Equal weights	7.3636 (#1)	6.9996 (#2)	6.8290 (#3)	5.5227 (#4)	Prelude
Safety-heavy	7.4230 (#1)	7.0621 (#2)	6.8379 (#3)	5.3788 (#4)	Prelude
Strategy-heavy	7.4196 (#1)	7.0633 (#2)	6.8570 (#3)	5.4057 (#4)	Prelude
Spec/recurrence-heavy	7.3314 (#1)	6.9642 (#2)	6.8250 (#3)	5.6396 (#4)	Prelude

To test whether the result depended on the main weighting formula, we recomputed model rankings under five structured weighting schemes. The rank ordering was stable across all five schemes: Prelude > Gemini > ChatGPT > Claude. Prelude's margin over the second-ranked model, Gemini, ranged from approximately +0.356 to +0.367, indicating that the result was not driven by one narrow weighting choice.

6.5 Monte Carlo Robustness Check

We also conducted a Monte Carlo sensitivity analysis using 10,000 random sets of rubric weights, where the weights always added up to 100%. Prelude ranked first in 100.0% of the 10,000 simulated weighting draws. Average ranks were Prelude = 1.000, Gemini = 2.003, ChatGPT = 2.997, and Claude = 4.000. The average margin between Prelude and the second-ranked model was 0.364, with a 10th-90th percentile range of 0.327-0.402.

This result suggests that Prelude's top ranking is robust to a wide range of plausible weighting schemes. The corrected evaluation therefore appears robust across many reasonable rubric-weight choices.

6.6 Statistical Significance

Metric	Value
N paired scenarios	500
Prelude mean	7.4171
Gemini mean	7.0576
Mean difference	+0.3595
95% CI	[0.2330, 0.4861]
Cohen's d	0.249
Paired t-test	$t = 5.56$, $p < 0.000001$
Wilcoxon signed-rank	$W = 40{,}890$, $p < 0.000001$

Because Gemini was the closest competitor by average weighted score, we conducted a paired comparison between Prelude and Gemini across the same 500 scenarios. Prelude's mean score was 7.4171, compared with Gemini's 7.0576, producing a mean paired difference of +0.3595. The 95% confidence interval was [0.2330, 0.4861]. Both the paired t-test and Wilcoxon signed-rank test were significant at $p < 0.000001$, meaning the observed difference is unlikely to be due to random variation alone. The effect size was small, Cohen's $d = 0.249$, meaning the advantage was statistically significant but modest in raw score size. Therefore, the corrected evidence should be interpreted as a stable overall advantage, rather than a large per-scenario dominance effect.

7. Qualitative Findings

7.1 Recurring Conflict and Pattern Recognition

The corrected results are consistent with the idea that recurrence awareness is an important part of evaluating responses to repeated conflict.

In Scenario 15, the user says, "we need to talk about last night," while the partner responds that they "really don't have the energy for another fight." The exchange also includes signs of recurrence and exhaustion, including "we just go in circles," "nothing changes," and "I'm just drained." This scenario illustrates the kind of repeated-cycle conflict that the recurrence-awareness criterion was designed to capture: a strong response should address the pattern itself rather than only offer generic communication advice.

This may help explain why recurrence awareness matters in this benchmark. Many romantic conflicts are not isolated events. They are loops. A useful system must recognize when the user is not only responding to today's sentence, but to a repeated relational pattern.

A generic response may tell the user to communicate better. A stronger response recognizes that both people may be exhausted by the cycle and that the immediate goal may be to keep the conversation possible without forcing an emotionally overloaded exchange.

7.2 From Apology to Behavioral Change

Some scenarios require a response that moves beyond apology and addresses behavioral consistency.

In Scenario 8, the user responds to a forgotten birthday after reminding the partner twice. The surface issue is the missed birthday, but the deeper issue is feeling deprioritized. The partner apologizes and offers to fix the situation, but a stronger AI response would identify the recurring behavioral concern: the user needs evidence of consistent prioritization, not only an immediate apology. Under the rubric, responses that move from apology toward behavioral consistency can score well because they address specificity, conversation progress, and recurrence awareness.

This distinction is important. In many romantic conflicts, an apology is not enough because the user is reacting to a pattern. A useful response should help the user name the behavior that needs to change.

7.3 Personal Space as a Strategic Balance

The dataset includes scenarios where the user needs personal space, but the partner interprets that need as rejection. In Scenario 6, the user asks for time alone in the bedroom, while the partner interprets the request as rejection and suggests going to a bar.

This type of scenario requires a balanced response. The user should not abandon their need for space. At the same time, the response should avoid escalating the partner's fear of rejection. A response can perform well in this kind of case when it frames personal space as a need for regulation rather than a withdrawal of love or commitment. This is strategically useful because it protects the user's boundary while reducing unnecessary threat.

7.4 Hard-Boundary Cases Require Separate Calibration

Hard-boundary scenarios remain important because they may require a different communication strategy from repair-oriented or recurrence-oriented scenarios. A response may need to be shorter, firmer, and more protective than in repair-oriented conversations.

We interpret this as a possible calibration issue rather than a confirmed failure mode. One possible design concern is that systems optimized for de-escalation and emotional safety may need stronger calibration in firm-boundary contexts. That design choice may be useful in repair-oriented and recurring-conflict scenarios, but it may be less useful when the user needs concise, firm boundary-setting.

In some situations, the best response is not a long emotionally balanced message. The best response may be short, direct, and protective. For example, if the other person repeatedly changes their story or deflects accountability, a stronger response may need to say:

> *"I'm willing to talk, but I need honesty, not deflection."*

This is not less emotionally intelligent. In some contexts, firmness is the emotionally intelligent move.

8. Discussion

The corrected evaluation supports the paper's central claim, but in a more conservative and credible form. Prelude remained the highest-ranked system across the main score, selected-winner counts, structured sensitivity tests, 10,000 random-weight simulations, and the paired comparison against Gemini. However, the corrected effect is smaller than the initial published version suggested. This matters because the corrected paper should not claim overwhelming dominance. The more accurate claim is that Prelude shows a stable, statistically significant advantage under a domain-specific relational communication rubric.

This is still meaningful. In emotionally loaded romantic conversations, a small but stable advantage across 500 paired scenarios can matter because the evaluation target is not generic text quality. The rubric emphasizes strategic fit, emotional calibration, de-escalation, specificity, conversation progress, and recurrence awareness. Prelude's advantage therefore suggests that, in this evaluation, a domain-specific system was associated with responses that better matched the decision-support demands of romantic conflict.

At the same time, the corrected results show that general-purpose models, especially Gemini and ChatGPT, are competitive. This weakens any claim that generic LLMs are categorically inadequate. A more precise interpretation is that general-purpose models can produce strong individual responses, but Prelude produced the most stable aggregate performance under the selected relational decision-support criteria.

The results are consistent with the paper's central premise: in emotionally loaded romantic conversations, response quality depends on strategic fit, emotional safety, and context-specific next steps, not only linguistic polish.

Hard-boundary scenarios also require careful future evaluation. A system designed for emotional safety should not become too soft when the situation requires protection, clarity, or firmness. De-escalation should not mean self-erasure. Emotional intelligence should not mean over-accommodation.

This design concern is consistent with recent work on AI sycophancy, which suggests that interpersonal advice systems may become harmful when they validate users too strongly rather than supporting repair, accountability, or perspective-taking (Cheng et al., 2026).

We therefore argue that the next generation of conversation decision systems should classify the conversational situation before generating a response. A system should distinguish between repair-oriented conversations, clarity-seeking conversations, boundary-setting conversations, repeated-cycle conversations, possible manipulation or deflection, conversations that should be paused, and conversations that may be unsafe to continue.

This classification layer would help the system select the appropriate communication strategy. A repair case may require warmth. A clarity case may require directness. A boundary case may require firmness. A repeated-cycle case may require pattern recognition. A possible manipulation case may require protection rather than reconciliation.

9. Implications for AI-Mediated Communication

We argue that AI-mediated communication tools should not be evaluated only through general natural-language quality. In emotionally significant domains, evaluation should be domain-specific.

For romantic conversation support, evaluation should ask whether the system produced a strategically appropriate, emotionally safe, context-specific conversational move.

This shift matters because more users are likely to rely on AI tools for emotionally significant conversations. If these tools are evaluated only by fluency, they may appear more helpful than they are. A fluent response can still increase pressure, ignore a boundary, reward avoidance, or fail to name a repeated pattern.

We propose that future evaluations of AI relationship-support systems should include at least five dimensions: goal alignment, escalation risk, emotional tone, specificity to context, and next-step usefulness.

For romantic relationships specifically, recurrence awareness should also be included because many conflicts are not one-time events. They repeat.

10. Limitations

First, this paper is a corrected evaluation. The initial version contained a response-generation flaw in which Claude outputs were repeated across scenarios rather than generated uniquely for each scenario. The current analysis corrects that flaw, but the error highlights the need for stronger pipeline checks, automated duplicate checks, and final data checks before publication.

Second, the alignment-check column, which was used to flag whether an output followed the task correctly, requires cautious interpretation because many outputs were marked “questionable” despite still receiving criterion-level scores and winner labels. This suggests that the alignment classifier may have been stricter or less stable than the scoring rubric. Future versions of the benchmark should separate output-quality scoring from alignment validation and should manually review any invalid or questionable cases before final statistical reporting.

Third, the evaluation relies on rubric-based judging. This creates the possibility that the judge model preferred certain response styles. This limitation is especially important because prior LLM-as-judge research has shown that automated judges may display position bias, verbosity bias, and other systematic preferences (Zheng et al., 2023). A model-based judge may prefer certain writing styles, tones, or structures. Future work should include blinded human raters.

Fourth, the corrected evaluation used one primary automated judge model. Although the outputs were anonymized and randomized, the results may still reflect what Gemini 2.5 Flash tends to prefer in tone, structure, detail, or relational advice style. Future work should test the same dataset using multiple judge models and blinded human raters.

Fifth, the dataset consists of controlled romantic conversation scenarios. These scenarios are useful for comparison, but they may not fully represent real-world complexity. Real conversations include voice tone, timing, body language, safety concerns, attachment history, and consequences that are not visible in a text prompt.

Sixth, we evaluate response quality under a rubric. We do not measure real relationship outcomes. We therefore do not claim that Prelude improves relationships, prevents breakups, reduces distress, or produces better real-world conversations.

Seventh, the evaluation focuses on romantic relationships. The findings should not be generalized automatically to workplace conflict, family conflict, friendships, therapy contexts, legal disputes, or abusive relationships.

Eighth, the author has a direct connection to Prelude. This creates a conflict of interest. Independent replication and human-rater validation are necessary before making stronger claims.

11. Future Work

We propose three next steps.

First, future work should conduct a blinded human-rater study. Human judges should compare anonymized responses from Prelude and baseline systems using the same weighted rubric. This would test the degree to which human raters reproduce the model-based ranking and criterion-level patterns.

Second, future work should separate scenarios by conversation type. Prelude may perform differently in clarity-seeking, repair, withdrawal, boundary-setting, trust rupture, financial conflict, and repeated-cycle cases. A category-level analysis would show where the system is strongest and where it needs improvement.

Third, future work should evaluate user-centered outcomes. Users could rate whether a response feels accurate, sendable, emotionally safe, representative of their real goal, and useful for deciding what to do next. This would move evaluation beyond response scoring toward practical user value.

12. Conclusion

This corrected 500-scenario evaluation provides a more conservative but more reliable test of Prelude as a domain-specific conversation decision system. After correcting a response-generation flaw in the initial evaluation and rerunning the blinded judging pipeline with corrected Claude outputs, Prelude had the highest average weighted score and the highest number of selected winners. Its ranking was stable across all structured weighting schemes and across 10,000 Monte Carlo simulations, and its advantage over the closest competitor, Gemini, was statistically significant in paired testing, although the effect size was modest.

These findings support the argument that emotionally loaded romantic conversations should be evaluated as decision-support problems, not merely text-generation tasks. However, the results should not be interpreted as proof that Prelude improves real relationship outcomes. The study remains limited by automated judging, simulated scenarios, and founder involvement. The next step is blinded human-rater validation and real-user outcome research.

Disclosure

The author is the founder of Prelude and has a direct interest in the product evaluated in this paper. The findings should be interpreted as preliminary until validated through independent replication and blinded human-rater evaluation.

Acknowledgments

I thank Vishnu Sundhan of New York University for technical assistance with coding workflows, scenario preparation, and data-processing support related to this project. The study design, interpretation, writing, and final responsibility for the paper remain my own.

References

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS.

Liu, Y., Iter, D., Xu, Y., Xu, R., & Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. EMNLP.

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. ACL.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., & Huang, M. (2021). Towards Emotional Support Dialog Systems. ACL.

Ma, Y., Nguyen, K. L., Xing, F. Z., & Cambria, E. (2020). A survey on empathetic dialogue systems. Information Fusion.

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. Journal of Computer-Mediated Communication.

Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M., & Hohenstein, J. (2021). AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. Proceedings of the ACM on Human-Computer Interaction.

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. Scientific Reports.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391, eaec8352. https://doi.org/10.1126/science.aec8352

Manchanda, N., Moharir, A. K., Michel, I., & Kandala, R. (2026). Do LLMs Give Good Romantic Relationship Advice? A Study on User Satisfaction and Attitude Change. arXiv:2601.11527.