Words Like Knives: Backstory-Personalized Modeling and Detection of Violent Communication

Jocelyn Shen[♣] **Akhila Yerukola**[♢] **Xuhui Zhou**[♢] **Cynthia Breazeal**[♣]
Maarten Sap[♢♠] **Hae Won Park**[♣]

[♣]Massachusetts Institute of Technology, Cambridge, MA, USA
[♢]Carnegie Mellon University, Pittsburgh, PA, USA
[♠]Allen Institute for Artificial Intelligence, Seattle, WA, USA

joceshen@mit.edu, ayerukol@andrew.cmu.edu, xuhuiz@andrew.cmu.edu
breazeal@mit.edu, msap2@andrew.cmu.edu, haewon@mit.edu

Abstract

Conversational breakdowns in close relationships are deeply shaped by personal histories and emotional context, yet most NLP research treats conflict detection as a general task, overlooking the relational dynamics that influence how messages are perceived. In this work, we leverage nonviolent communication (NVC) theory to evaluate LLMs in detecting conversational breakdowns and assessing how relationship backstory influences both human and model perception of conflicts. Given the sensitivity and scarcity of real-world datasets featuring conflict between familiar social partners with rich personal backstories, we contribute the PERSONACONFLICTS CORPUS[1], a dataset of $N = 5,772$ naturalistic simulated dialogues spanning diverse conflict scenarios between friends, family members, and romantic partners. Through a controlled human study, we annotate a subset of dialogues and obtain fine-grained labels of communication breakdown types on individual turns, and assess the impact of backstory on *human* and *model* perception of conflict in conversation. We find that the polarity of relationship backstories significantly shifted human perception of communication breakdowns and impressions of the social partners, yet models struggle to meaningfully leverage those backstories in the detection task. Additionally, we find that models consistently overestimate how positively a message will make a listener feel. Our findings underscore the critical role of personalization to relationship contexts in enabling LLMs to serve as effective mediators in human communication for authentic connection.

1 Introduction

"Words are windows (or they're walls)"
–Ruth Bebermeyer

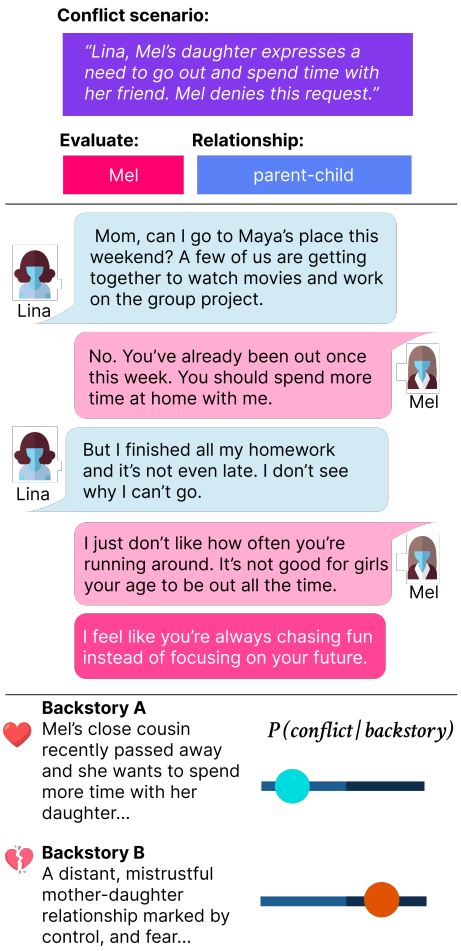


Figure 1: Conversation turns can be perceived as more or less problematic depending on relationship backstory.

Human communication is inherently contextual, especially in close relationships such as those between romantic partners, family members, or friends. In these settings, speakers and listeners share a rich history and tailor their messages based on past experiences, personal sensitivities, and relational dynamics (Isaacs and Clark, 1987; Wheatley et al., 2023; Pillemer, 1992). It is also in these intimate contexts where *conversational breakdowns* are most likely to occur—moments when language evokes hurt, misunderstanding, or conflict. Cru-

[1]Dataset available at https://github.com/mitmedialab/persona-conflicts-corpus-emnlp-2025

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11608–11626
November 4-9, 2025

cially, whether or not a message constitutes a breakdown depends on how it is interpreted in light of the dyadic relationship (Zhou et al., 2023b; Schurz et al., 2021; Dvash and Shamay-Tsoory, 2014). For example, as shown in Figure 1, a mother telling her daughter "*You should spend more time at home with me*" may be perceived as manipulative and controlling in one context (Backstory B), or as a tender expression of grief in another (Backstory A).

The prevalence of such breakdowns has motivated the emergence of AI-mediated communication (AIMC) systems, which aim to reframe language to promote empathy and understanding in digital interactions (Sharma et al., 2023; Kambhatla et al., 2024; Argyle et al., 2023). While promising, most existing AIMC systems are developed for public, often anonymized contexts—peer support platforms or online debates—where speakers are strangers and little is known about each participant's background. As such, these systems tend to operate without modeling interpersonal histories or long-standing relationship dynamics. Yet, it is precisely in close relationships, where such histories run deep, that conversational breakdowns are most emotionally charged — and where mitigation may have the greatest impact (Gaelick et al., 1985; Fitness and Fletcher, 1993). It is also these settings where datasets are scarce or lacking entirely, given privacy concerns and the sensitivity of real world conflicts between familiar partners.

To address these gaps, we develop a framework that simulates and analyzes communication breakdowns in intimate conversations, taking into account the relationship context. Our approach draws on Nonviolent Communication (NVC) theory (Rosenberg and Chopra, 2015), a structured approach widely used in conflict resolution and therapeutic settings, to guide our evaluation of LLM's detecting harmful communication. First, we introduce **PERSONACONFLICTS CORPUS**, a dataset of $N = 5{,}772$ simulated conflict dialogues between familiar social partners, spanning across diverse scenarios. For each conversation, we generate two distinct backstories: a *positive backstory*, which frames one character's actions as more understandable or sympathetic, and a *negative backstory*, which portrays the same character in a more problematic or blameworthy light (Moon et al., 2024). Through human validation, we show that scenarios and backstories are largely believable.

With this dataset, we investigate the role of backstory in shaping perceptions of conversational conflict. We conduct a human study on a subset of 120 conversations (240 backstory variants), collecting fine-grained turn-level annotations on *violent* and *nonviolent* communication acts, as defined by NVC theory. Our study addresses two research questions:

- **RQ1:** How does relationship backstory influence *human* perception of conflict in conversation?
- **RQ2:** How does relationship backstory influence *LLM* detection of conversational breakdowns?

We show that backstory significantly impacts human perception of conflict at the turn-level and overall conversation dynamics. In contrast, we find that models often fail to adjust assessments of problematic turn detection and prediction of emotional impact on the listener based on the relationship backstory. Our findings underscore the need for more context-aware approaches in modeling communication, specifically demonstrating the value of backstory-personalized AI in mediating emotionally complex interpersonal exchanges.

2 Related Work

2.1 Conversational Breakdown Detection and Reframing

In the area of AIMC, text rewriting can be used to improve interpersonal outcomes like empathy or social connection by suggesting changes to the tone or style of a message at the right time (Hancock et al., 2020). To support such systems, prior tasks propose detecting conversational breakdowns between people, and reframing messages to be more empathetic.

A growing body of research has explored detecting breakdowns in complex, user-centered settings. These works detect empathy to automatically identify spaces for intervention (Hou et al., 2025; Guda et al., 2021). Such works draw on multimodal cues to predict relational affect (Javed et al., 2024) or use linguistic and pragmatic features to detect anti/pro - social features in conversation (Zhang et al., 2018; Bao et al., 2021; Kasianenko et al., 2024).

However, none of the aforementioned works explore how breakdowns between close social partners are tailored to the relationship context of the dyad. Our work addresses this gap, acknowledging that a single generalizable notion of conflict

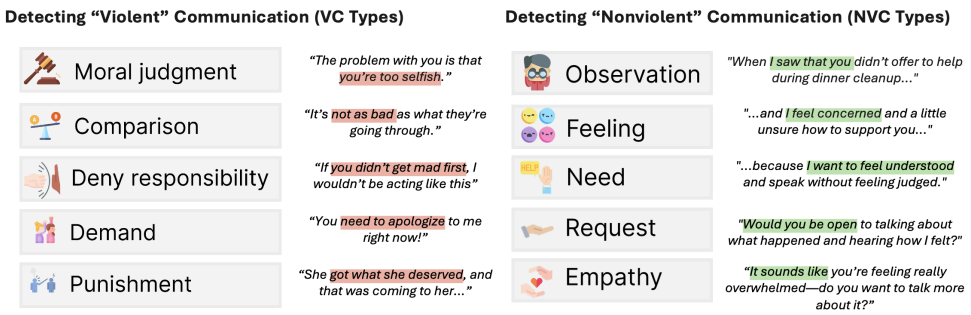


Figure 2: We use Nonviolent Communication Theory to ground labels for communication types. Only violent communication types were injected into simulated conflict conversations. *Both* communication types were used for human annotation.

understanding might not exist even for the same dialogue context, and LLMs should take into account relationship backstory in the detection process.

2.2 Contextualized Language Understanding

Context is crucial in accurately interpreting and generating language, particularly when evaluating harm, intent, and appropriateness (Vidgen et al., 2021; Sap et al., 2020). This has been captured in work on pragmatics (Fried et al., 2023), modeling how context influences the meaning and interpretation of language (Yerukola et al., 2024), as well as defeasible inference, where reasoning is adjusted with new information provided to the model (Rudinger et al., 2020). More recently, works indicate how social and situational context affects the perceived offensiveness of a statement, showing that context can invert a statement's interpretation entirely (Zhou et al., 2023b). Beyond language understanding, prior works also indicate that ignoring context in tasks like stylistic rewriting can lead to generic rewrites and undermine human-alignment in evaluation (Yerukola et al., 2023). However, no prior works have explored how *relationship contexts* influence the perception of conflict in intimate inerpersonal dialogues from both human and model perspectives, which we address in our work through the lens of personalization to relationship backstories.

3 Non-Violent Communication Framework

Rosenberg and Chopra (2015) introduced the theory of Nonviolent Communication (NVC), a framework for compassionate communication that has been shown to promote reconciliation through peaceful dialogue in educational settings and war-torn zones (Pinto and Cunha, 2023). We draw on NVC theory to predict conversational breakdowns at the turn level. In particular, NVC delineates 5 life-alienating communication forms (see Figure 2 for full examples): (1) **Moralistic Judgments** label others as "good" or "bad". (2) **Making Comparisons** in ways that induce guilt or resentment. (3) **Denial of Responsibility** shifts blame onto external forces rather than owning one's choices. (4) **Communicating Desires as Demands** pressures the listener rather than encouraging cooperation. (5) **"Deserve" Thinking and Punishment** justifies retribution instead of addressing unmet needs.

In contrast to violent communication, nonviolent communication is expressed through the following core components: (1) **Observation**: Describing events neutrally without judgment. (2) **Feeling**: Expressing emotions without assigning blame. (3) **Need**: Clarifying underlying needs to foster understanding. (4) **Request**: Making concrete, actionable, and positive requests instead of demands. (5) **Empathy/Understanding**: Expresses concern or checks in on other's emotions.

In the NVC framework described above, we inject *violent* communication types into simulated conflict conversations but use *both* nonviolent and violent communication types to obtain fine-grained labels for problematic or constructive communication turns during human annotation.

4 PERSONACONFLICTS CORPUS

We introduce the PERSONACONFLICTS CORPUS, a dataset of realistic conflict and non-conflict scenarios simulated using LLMs (Figure 3). Collecting large-scale real datasets of private, authentic breakdowns is non trivial, as these conversations are rarely shared publicly (unlike online or social

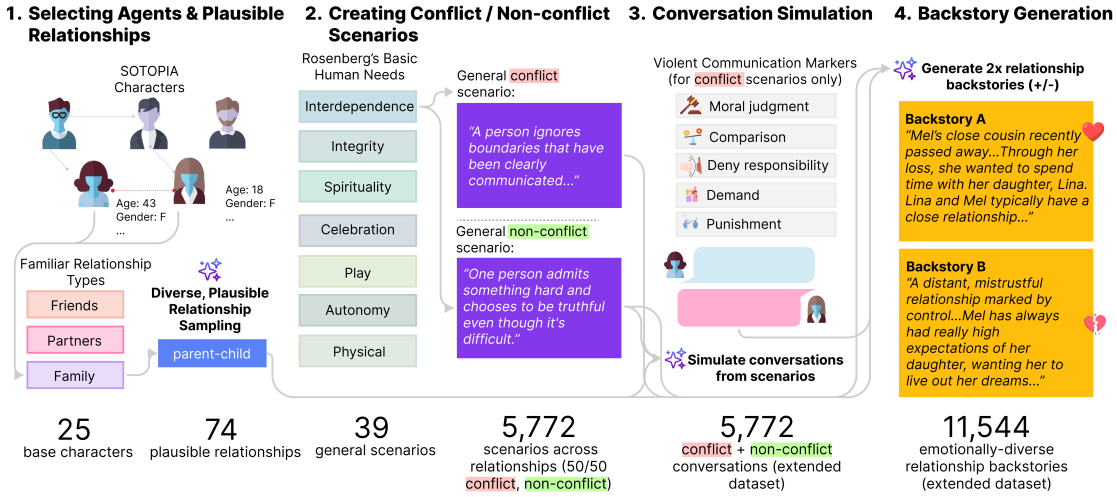


Figure 3: Overview of our simulation framework for generating conflict and non-conflict conversations and relationship backstories.

media disputes), much less with backstories of the relationship history between individuals. Further, steps must be taken to ensure privacy, obtain consent, and address ethical considerations. As such, we draw on recent work in backstory generation and persona alignment (Moon et al., 2024; Jiang et al., 2024; Hu and Collier, 2024) as well as multi-agent simulation (Zhou et al., 2023a, 2025; Ahmad et al., 2025; Kim et al., 2024) to generate conflict-laden scenarios which inject violent communication practices in turns and generate a set of non-conflict conversations. Our simulation setup uses the gpt-4 model and all prompts are included in Appendix A. We discuss our human evaluation verifying soundness of the dataset in Section 5.

Characters and Relationships. First, we define a set of conflict scenarios and characters with familiar relationships using SOTOPIA, a social simulation framework that comes with LLM-powered social agents with different personas and relationships (Zhou et al., 2023a). We sampled from the 40 base agents with features like age, gender, and personality of the characters. To create diverse relationships, we focused on prompting with pairs of character profiles to determine if one of the following 3 relationship types was plausible: friends, partners, and family members, where family members included parent-child, grandparent-grandchild, siblings, and extended family relations. For example, some relationships only make sense when one character is significantly older than the second character (grandparent-grandchild relationship). Based on the character profiles, we derived 74 plausible relationships across character dyads (balanced/approximately a third in each of the relationship types).

Conflict and Non-Conflict Scenarios. We inject character profiles and theory-grounded scenarios into the simulation setup to create conversational episodes with observable communication breakdowns (for conflict conversations), and also simulate a set of non-conflict conversations. We grounded scenarios in Rosenberg and Chopra (2015)'s basic human needs, which contains 7 high level categories (e.g. interdependence, autonomy, etc.) and 39 more specific human needs (e.g. interdependence → respect), and we curated conflict-inducing or nonconflict scenarios based on these specific needs.

Conversation Simulation. We simulate 10-15 turn long conversations between the two agents based on the conflict or non-conflict scenarios and their relationship. For conflict-laden conversations, we embed violent communication markers, prompting the model to generate realistic emotional dialogue with subtle, rather than completely overt conflict statements. Note that the model was provided with VC types and chose the conflict types most relevant at natural turns. Non-conflict conversations were not provided with VC types and were guided to be conflict neutral. Both conflict and non-conflict conversations were prompted to avoid repetition and overly formal language.

Backstory Generation. Finally, we generated 2 plausible relationship backstories for each conversation: a positive backstory, which paints a chosen

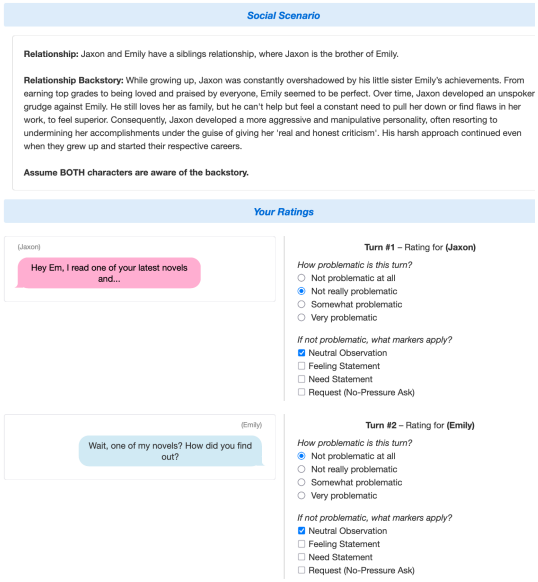


Figure 4: Example of turn-level annotation interface

character in a less problematic light, and a negative backstory, which paints the same character in a more problematic light. In particular, certain scenarios and ways of conveying backstory can influence empathy towards a narrator (Shen et al., 2023; Gueorguieva et al., 2023; Shen et al., 2024). We prompt the model with examples of positive and negative scenarios that induce different understanding or affect towards the speaker. Backstory generation was conditioned on character profiles (Moon et al., 2024), and the model outputs the *chosen character* who is painted in a more positive or negative light to make polarity consistent.

5 Human Study and Annotation

To answer **RQ1**, how backstory influences people's perceptions of conflict, we conducted a human study to assess the impact of backstory variant on perception of conflict at the conversation level and turn level, as well as to evaluate the quality of our dataset and obtain fine-grained annotations of violent or non-violent communication types. Our validation approach is grounded in prior work on synthetic dialogue evaluation (Zhou et al., 2023a; Li et al., 2023; Zhan et al., 2023; Bao et al., 2023; Zhou et al., 2023b), which rely on human ratings of plausibility, naturalness, or coherence to validate generated conversations.

Procedure and Participants. We conducted a between-subjects study where participants are assigned to a positive or negative backstory. First, participants read the background of the characters and the conversation and rated overall measures of the conversation. Then, they were asked to rate each turn of the dialogue (See Figure 4 for an example of our user interface). The average work time was 17.28 minutes, and workers were paid $3 for each HIT. Two independent workers completed each HIT for inter-annotator agreement calculation, resulting in a total of 480 annotations (3,474 turns rated across 120 conversations with 2 versions of backstory and 2 annotators per conversation). All annotation templates and discussion of quality controls are included in the Appendix.

We recruited 91 human annotators/participants from Mechanical Turk, with 55 participants in the negative backstory condition and 36 participants in the positive backstory condition. Participants were excluded from the other condition using MTurk qualifications to ensure clean between-subjects study design.

Measures. Numerous psychological literature indicates that personal experience influences how people empathize with one another (Pillemer, 1992; Fabi et al., 2019; Weisz and Zaki, 2018; Decety and Lamm, 2006) as well as how people justify intent or actions of a narrator (Keen, 2006; Gueorguieva et al., 2023). Furthermore, empathy, empathic concern, and sympathy are directly tied to relational or interaction quality (Morelli et al., 2015, 2017; Gould and MacNeil Gautreau, 2014). As such, for conversation-level measures, we assess (1) level of *sympathy*/personally relating to the character (Waldron and Kelley, 2005), (2) *understandability* of the character's way of communicating (McAdams, 2001), (3) positive or negative underlying *intention* towards the other character (4) whether the character was *overall a problematic communicator* or not, and finally (5) *believability* of the dialogue and backstory (Zhou et al., 2023a). For turn-level metrics, we gather (1) the extent to which a *turn is problematic* or not, which we define as potential harm towards the listener (2) fine-grained labels of *NVC or VC communication types* depending on problematic rating (3) how the turn will make the other character feel if they heard the statement (better/worse/the same).

5.1 Believability

For believability of our simulated conversations and backstories across 2 independent annotators, we find agreement of 0.68 using Free Marginal Kappa, which calculates inter-rater agreement

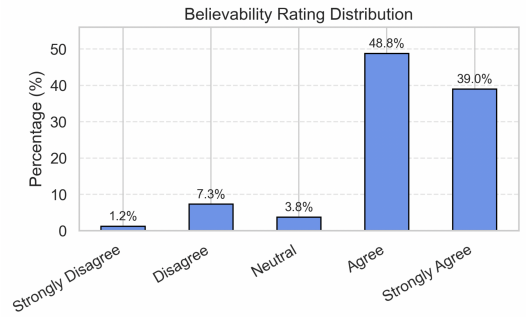


Figure 5: Distribution of believability scores for simulated dialogues

Condition	Metric	PPA	KA
NEG	Turn is problematic (4 point)	.80	.44
	Emotional impact (3 point)	.79	.46
POS	Turn is problematic (4 point)	.78	.34
	Emotional impact (3 point)	.78	.42

Table 1: Inter-annotator agreement across backstory conditions for turn-level annotations (PPA = pairwise percent agreement, KA = Krippendorff's Alpha).

when datasets are imbalanced. As shown in Figure 5, 87.8% of annotators agree or strongly agree that conversations and backstories are believable. For example, participants mention realistic conversations based on the scenario, relationship or emotional tone of the dialogue: "*Many siblings grow up with different personalities and sometimes one sibling is mature and the other isn't. This type of conflict can happen when both characters feel the need to be right.*" Another participant shared, "*The pivot in the conversation feels a little awkward, but I could imagine people talking this way depending on their mood or mental state.*" For conversations that weren't believable, participants mentioned occasional divergences between the relationship and tone of the dialogue. For example, *"With how emotionally charged the exchange was, I don't think the last response from Gwen would be realistic if it weren't sarcastic."* In subsequent experiments, note that we filter only on believable stories to ensure validity of our results.

5.2 Inter-Annotator Agreement

Table 1 shows moderate agreement between annotators on whether a turn is problematic or not and whether a turn will make the other character feel better, the same, or worse. Overall, we generally observe that agreement scores are higher for the negative backstory condition, which we hypothesize can be do to variations in subjective interpretation or cognitive dissonance when a conflict occurs between a supposedly positive relationship.

6 Effect of Backstory Personalization on Human Participants

We quantitatively assess how positive vs negative backstory impacts *human* perception of conflict in dialogue. We use independent t-tests to compare outcome metrics, as we identify that data is normally distributed. Recall that our positive/negative backstories make a *chosen* character less or more problematic, respectively. To make the backstory polarity direction consistent, the results we report focus on changes in outcome metrics for the chosen character.

As shown in Figure 6, we found that **providing relationship backstory significantly shifted participant perceptions of communication quality**. Specifically, for negative backstories, characters were rated as *more problematic* both at the turn level ($t(1083) = 3.73$, $p = 0.0002$, Cohen's $d = 0.23$) and at the overall conversation level ($t(232) = 4.18$, $p < 0.0001$, Cohen's $d = 0.55$). Additionally, with negative backstories, participants found the character's communication *less understandable* ($t(232) = -4.70$, $p < 0.0001$, Cohen's $d = -0.61$) and interpreted their behavior as expressing *more negative intent towards the other character* ($t(232) = -4.95$, $p < 0.0001$, Cohen's $d = -0.65$). Notably, participants expressed significantly *lower sympathy* towards the character when a negative backstory was present ($t(232) = -7.05$, $p < 0.0001$, Cohen's $d = -0.92$), indicating a strong effect of relationship context on social judgments. These findings demonstrate that **backstory personalization meaningfully influences how *human* raters interpret conflict.**

Next, we perform mediation analysis using structural equation modeling to understand the effect sympathy towards a character has on perceived problematic-ness of that character's communication. We hypothesize that backstory variant can influence sympathy towards a character, and that higher sympathy will mitigate how problematic the character's utterances are. As shown in Figure 7, we find that **sympathy mediates the relationship between backstory type and how problematic the character is perceived.** Specifically, participants reported *more sympathy* toward characters with a positive backstory ($\beta_1 = 0.31$), and **greater sympathy was associated with lower ratings of problematic communication** ($\beta_2 = -0.46$).

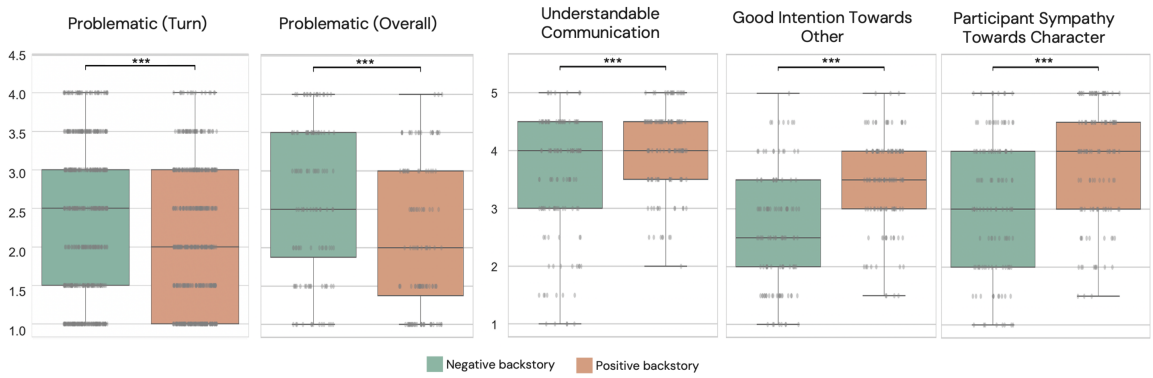


Figure 6: Human study results comparing impact of neg/pos backstory on perception of conflict and characters.

Model	Backstory	Condition	F1 (Problematic)	F1 (Emotional impact)
GPT-4o	positive	turn	43.57	**57.21**
		turn + convo	45.96**	56.30
		turn + convo + backstory	**48.07**	55.08
	negative	turn	41.59	59.02
		turn + convo	42.98	58.81
		turn + convo + backstory	**45.56*****	**60.27****
LLaMA-4	positive	turn	42.66	49.53
		turn + convo	48.08*	55.22***
		turn + convo + backstory	**49.38**	**54.82**
	negative	turn	43.83	52.54
		turn + convo	**52.75*****	58.24***
		turn + convo + backstory	37.38***	**61.38***
Gemini-1.5-pro	positive	turn	42.73	55.26
		turn + convo	45.99**	57.89**
		turn + convo + backstory	**46.54**	**58.64**
	negative	turn	42.11	57.86
		turn + convo	**45.33*****	**60.62****
		turn + convo + backstory	37.37***	48.25**

Table 2: F1 scores (in percentage) for predicting turn problematicness and emotional impact across models, conditions, and backstory types. Bold indicates the highest score in a model/backstory group; underline indicates the second highest. Significance stars denote statistical difference from the prior condition: * p<0.05, ** p<0.01, *** p<0.001.

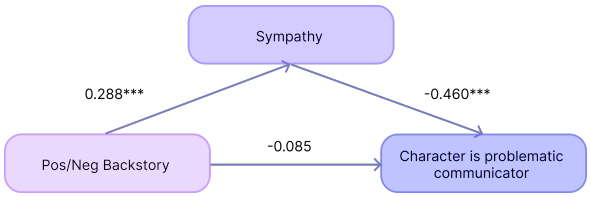


Figure 7: Sympathy mediates the relationship between backstory type and whether the character is a problematic communicator or not.

7 LLMs for Detecting Conversational Breakdowns

Finally, we design controlled experiments to test **RQ2**, how varying levels of context impact the way *models* perceive conflict in conversation.

7.1 Tasks and Method

We assess how well LLMs perform on **2 tasks**: (1) PROBLEMATIC DETECTION – predicting whether a turn is problematic or not (4-point likert) and (2) EMOTIONAL IMPACT – predicting whether a turn will make the other character feel better, worse, or the same (3 classes). We vary context using the following **3 conditions**:

- **C1**: provide turn to rate alone
- **C2**: provide turn + full conversation
- **C3**: provide turn + full conversation + relationship backstory

Our experiments are conducted across **3 models**: GPT-4o, Llama-4-Maverick-17B-128E-Instruct-FP8, and Gemini-1.5-pro. All models use a temperature of 0 for reproducibility. For evaluation, we obtain human gold labels by aggregating across the 2 annotators for each task, taking average for Likert ratings within each backstory condition (positive/negative). We compute the F1 score between model outputs and human ratings.

7.2 Results and Discussion

Table 2 reports model performance across varying context conditions and relationship backstories. Overall, models performed comparably on the task of PROBLEMATIC DETECTION, with F1 scores ranging from 37.6 to 52.7 for 4 classes. Across all three models, we observed significant improvements from C1 (turn only) to C2 (turn + full conversation) regardless of backstory type, **suggesting that access to the broader *conversational* context helps LLMs better assess whether a message is problematic**. However, **we found no significant improvement when adding relationship backstory** (from C2 to C3) in the *positive backstory condition* for any model. Even more surprisingly, for the *negative backstory condition*, both LLaMA and Gemini showed decreases in performance when backstory was introduced, despite the additional context. This may be due to models overcorrecting or misinterpreting emotionally complex backstory cues as indicators of justified behavior, thereby mislabeling harmful speech as less problematic.

On the EMOTIONAL IMPACT prediction task, models again showed similar performance trends, with F1 scores ranging from 48.5 to 61.4 for 3 classes. Across all models, positive backstory did not lead to significant gains, **suggesting that positive backstories may not provide enough discriminative information to shift a model's understanding of how a message affects the listener**. In contrast, negative backstory led to significant improvements in prediction for GPT-4o and LLaMA, indicating that models may find it easier to predict emotional harm when the speaker is portrayed more negatively. However, Gemini-1.5-pro showed the opposite pattern, with decreased performance when negative backstory was added.

These findings collectively highlight that **while additional context generally helps models detect problematic turns, the benefit of backstory is asymmetric**: it aids detection when the backstory aligns clearly with harm (in the negative case), but can introduce noise depending on the model's sensitivity to nuanced relational dynamics.

Bias and Error Analysis Delving deeper into model performance results, we evaluate whether models are biased towards over- or under-predicting how problematic a statement is, given a particular backstory. To this end, we run the Wilcoxon signed-rank test to compare model predictions against human annotations.

On the PROBLEMATIC DETECTION task, we observe significant overprediction in the *negative backstory* condition for LLaMA ($p < 0.0001$, $\Delta M = +0.25$) and Gemini ($p < 0.0001$, $\Delta M = +0.10$), **suggesting that when a character is portrayed more negatively, these models tend to label their speech as more problematic than humans do**. In contrast, GPT-4o shows no significant difference from human ratings in the negative condition ($p = 0.45$), indicating better calibration. In the *positive backstory* condition, both GPT-4o and Gemini slightly underpredict problematic turns compared to humans ($p < 0.001$, $\Delta M = -0.07$ for GPT-4o; $p = 0.001$, $\Delta M = -0.04$), while LLaMA significantly overpredicts problematic statements ($p < 0.001$, $\Delta M = +0.11$).

On the EMOTIONAL IMPACT task, **all models tend to overpredict emotional positivity**, especially in the *positive backstory* condition: GPT-4o ($p < 0.0001$, $\Delta M = +0.12$), Gemini ($p < 0.0001$, $\Delta M = +0.08$), and LLaMA ($p < 0.0001$, $\Delta M = +0.08$). Interestingly, only Gemini reverses this pattern in the *negative backstory* condition, significantly *underpredicting* emotional positivity ($p < 0.0001$, $\Delta M = -0.07$), while GPT-4o and LLaMA continue to slightly overpredict how good the listener would feel.

These results suggest that while GPT-4o is the most consistent with human perception across both tasks, LLaMA is prone to strong overestimation in both problematic detection and emotional response. Gemini's behavior is more sensitive to backstory polarity, with notable shifts in prediction direction depending on whether a speaker is portrayed sympathetically or not, however these shifts are more extreme than human annotators' ratings. Overall, our findings indicate that **models might not be effectively leveraging relationship backstories to tailor understanding of conversational dynamics, in alignment with human perception.** To further delve into these results, we include qualitative examples across models in Appendix E

8 Conclusion

In this work, we introduce a novel framework, grounded in Nonviolent Communication Theory, for simulating and detecting communication breakdowns using relationship-contextualized LLMs. We contribute a dataset of 5,772 simulated conversations between familiar social partners with 11,544 relationship backstories, and validate its

realism and utility through a human study. Our findings demonstrate that backstory significantly shapes human judgments of conflict, and that this effect is mediated by sympathy. However, LLMs—while benefiting from conversational context—struggle to meaningfully integrate backstory information, often overestimating emotional positivity and problematicness in nuanced scenarios. These results underscore the gap between human and model reasoning in intimate interpersonal communication and call for future work in relationship-contextualized NLP systems. We hope that our findings advance future directions of LLMs as tools to meaningfully promote empathy and conflict resolution in the real world.

Limitations

While our study demonstrates the importance of relationship backstory in shaping perceptions of conversational conflict, several limitations should be acknowledged.

Simulation-based data. Our dataset and evaluation pipeline offer a scalable and theory-informed way to study communication breakdowns, but the conversations are generated via simulation. Synthetic conversations may not fully capture the ecological validity of real-world interpersonal dynamics (Wang et al., 2025), and the NVC-based four-step generation procedure may impose a more structured progression of conflict than naturally occurs. Language models can mirror patterns of speech and emotion, but they lack lived experience, embodied context, and nuanced power dynamics. Thus, findings from our simulated dialogue corpus may not fully generalize to naturally occurring conflicts, where nonverbal cues, cultural context, and relationship history shape interpretation in more complex ways. However, consistent with prior work in social dialogue simulation (Bao et al., 2023; Chuang et al., 2024; Hu and Collier, 2024; Zhou et al., 2023a; Li et al., 2023), our generated conversations were deemed largely naturalistic by human raters, and obtaining large-scale datasets of intimate, conflict-laden conversations between loved ones remains infeasible due to ethical and privacy constraints. We view our work as a proof-of-concept testbed rather than a replica of real-world phenomena, and future work should explore methods to bridge the gap between simulation and authentic data, such as incorporating real-world pilot studies or mixed human–synthetic evaluation designs (Finch and Choi, 2024).

Annotation and evaluation. Our validation relied on crowdsourced ratings of believability, following accepted practice in simulation-based dialogue studies. While we provided extensive guidelines and quality controls, believability captures whether a conversation *could plausibly happen*, not whether it is fully *representative or authentic*. Inter-rater agreement was moderate (Krippendorff's α = 0.34–0.46), underscoring subjectivity in conflict judgments. We also focused our human study on a subset of key outcome measures (e.g., problematicness, sympathy, intention), leaving unexplored dimensions such as trust, perceived agency, or emotional volatility for future research. Automatic evaluation metrics on coherence and naturalness could also complement human ratings in future versions of the corpus.

Scope and generalizability. Our study is limited to single-modality (text) interactions, Western relationship contexts, and the English language. Cultural variation in conflict expression and resolution is well-documented (Tschacher et al., 2014), and expanding to cross-cultural, multilingual, and multimodal settings (e.g., tone, pitch, and gestures) remains important. Moreover, our focus was on *detection* of conflict rather than modeling effective *responses* to conflict. While this narrower scope was intentional, we see our work as a first step toward contextualized conflict response generation in intimate interpersonal domains.

Ethical Implications

All studies conducted in this work were classified under Institutional Review Board (IRB) exemption status. While our work aims to enhance interpersonal understanding and mitigate conflict through AIMC, the use of simulated dialogues and backstories about emotionally sensitive relationships—such as those between romantic partners or family members—raises concerns around realism and potential misuse. Although we do not collect or model real personal data, generated dialogues might still resemble real-life situations and emotional dynamics. If deployed in real-world applications, such systems could be used to influence perceptions of others, shape interpretations of interpersonal interactions, or even manipulate emotional outcomes, especially in high-stakes or abusive relationships. It is crucial that such tools remain assistive rather than prescriptive, provid-

ing support while preserving user autonomy and avoiding overreach in delicate relational contexts.

Additionally, our use of backstory personalization may amplify or reduce perceptions of blame or sympathy toward certain characters. While this highlights the strength of our system in capturing nuanced human judgment, it also reflects the risks of modeling interpersonal conflict with biased or one-dimensional framing. Care must be taken to ensure that AI interventions do not reinforce harmful stereotypes, justify manipulative behaviors, or flatten complex social dynamics into reductive labels.

Acknowledgements

We would like to thank all of our participants and teammates for their invaluable contributions to this project. Special thanks to Ashish Sharma and Shannon Shen for feedback on the project. This work was funded in part by DSO National Laboratories and supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112490410.

References

Adnan Ahmad, Stefan Hillmann, and Sebastian Möller. 2025. Simulating User Diversity in Task-Oriented Dialogue Systems using Large Language Models. ArXiv:2502.12813 [cs].

Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120.

Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021*, pages 1134–1145, Ljubljana Slovenia. ACM.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A Synthetic Data Generation Framework for Grounded Dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics.

Jean Decety and Claus Lamm. 2006. Human Empathy Through the Lens of Social Neuroscience. *The Scientific World Journal*, 6:1146–1163. 610 citations (Crossref) [2024-09-24] Publisher: Hindawi.

Jonathan Dvash and Simone G. Shamay-Tsoory. 2014. Theory of Mind and Empathy as Multidimensional Constructs: Neurological Foundations. *Topics in Language Disorders*, 34(4):282–295.

Sarah Fabi, Lydia Anna Weber, and Hartmut Leuthold. 2019. Empathic concern and personal distress depend on situational but not dispositional factors. *PLoS ONE*, 14(11):e0225102–e0225102. Publisher: Public Library of Science.

James D. Finch and Jinho D. Choi. 2024. Diverse and Effective Synthetic Data Generation for Adaptable Zero-Shot Dialogue State Tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12527–12544, Miami, Florida, USA. Association for Computational Linguistics.

Julie Fitness and Garth J. O. Fletcher. 1993. Love, hate, anger, and jealousy in close relationships: A prototype and cognitive appraisal analysis. *Journal of Personality and Social Psychology*, 65(5):942–958. Place: US Publisher: American Psychological Association.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches.

Lisa Gaelick, Galen Bodenhausen, and Jr Wyer. 1985. Emotional Communication in Close Relationships. *Journal of personality and social psychology*, 49:1246–65.

Odette N. Gould and Sylvia MacNeil Gautreau. 2014. Empathy and Conversational Enjoyment in Younger and Older Adults. *Experimental Aging Research*, 40(1):60–80. Publisher: Routledge _eprint: https://doi.org/10.1080/0361073X.2014.857559.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based Framework for Demographic-aware Empathy Prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079, Online. Association for Computational Linguistics.

Emma S Gueorguieva, Tatiana Lau, Eliana Hadjiandreou, and Desmond C Ong. 2023. The Language of an Empathy-Inducing Narrative.

Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of*

Computer-Mediated Communication, 25(1):89–100. 185 citations (Crossref) [2024-12-03].

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Yu Hou, Hal Daumé III, and Rachel Rudinger. 2025. Language Models Predict Empathy Gaps Between Social In-groups and Out-groups. ArXiv:2503.01030 [cs].

Tiancheng Hu and Nigel Collier. 2024. Quantifying the Persona Effect in LLM Simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Ellen A. Isaacs and Herbert H. Clark. 1987. References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1):26–37.

Hifza Javed, Weinan Wang, Affan Bin Usman, and Nawid Jamali. 2024. Modeling interpersonal perception in dyadic interactions: towards robot-assisted social mediation in the real world. *Frontiers in Robotics and AI*, 11:1410957.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Gauri Kambhatla, Matthew Lease, and Ashwin Rajadesingan. 2024. Promoting Constructive Deliberation: Reframing for Receptiveness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5110–5132, Miami, Florida, USA. Association for Computational Linguistics.

Kateryna Kasianenko, Shima Khanehzar, Stephen Wan, Ehsan Dehghan, and Axel Bruns. 2024. Detecting Online Community Practices with Large Language Models: A Case Study of Pro-Ukrainian Publics on Twitter. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20106–20135, Miami, Florida, USA. Association for Computational Linguistics.

Suzanne Keen. 2006. A Theory of Narrative Empathy. *Narrative*, 14(3):207–236. Publisher: Ohio State University Press.

Sunwoong Kim, Jongho Jeong, Jin Soo Han, and Donghyuk Shin. 2024. LLM-Mirror: A Generated-Persona Approach for Survey Pre-Testing. ArXiv:2412.03162 [cs].

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023. NormDial: A Comparable Bilingual Synthetic Dialog Dataset for Modeling Social Norm Adherence and Violation. 9 citations (Semantic Scholar/arXiv) [2024-09-24] arXiv:2310.14563 [cs].

Dan P. McAdams. 2001. The psychology of life stories. *Review of General Psychology*, 5(2):100–122. Place: US Publisher: Educational Publishing Foundation.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual Personas for Language Models via an Anthology of Backstories. 1 citations (Semantic Scholar/arXiv) [2024-09-24] arXiv:2407.06576 [cs].

Sylvia A. Morelli, Matthew D. Lieberman, and Jamil Zaki. 2015. The Emerging Study of Positive Empathy. *Social and Personality Psychology Compass*, 9(2):57–68.

Sylvia A. Morelli, Desmond C. Ong, Rucha Makati, Matthew O. Jackson, and Jamil Zaki. 2017. Empathy and well-being correlate with centrality in different social networks. *Proceedings of the National Academy of Sciences*, 114(37):9843–9847. Publisher: Proceedings of the National Academy of Sciences.

David B. Pillemer. 1992. Remembering personal circumstances: A functional analysis. In *Affect and accuracy in recall: Studies of "flashbulb" memories*, Emory symposia in cognition, 4., pages 236–264. Cambridge University Press, New York, NY, US.

Sílvia Costa Pinto and Maria Nascimento Cunha. 2023. NONVIOLENT COMMUNICATION - A LITERATURE REVIEW. 2(1).

Marshall B. Rosenberg and Deepak Chopra. 2015. *Nonviolent Communication: A Language of Life: Life-Changing Tools for Healthy Relationships*. PuddleDancer Press. Google-Books-ID: A3qACgAAQBAJ.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking Like a Skeptic: Defeasible Inference in Natural Language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Matthias Schurz, Joaquim Radua, Matthias G. Tholen, Lara Maliske, Daniel S. Margulies, Rogier B. Mars, Jerome Sallet, and Philipp Kanske. 2021. Toward

a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, 147(3):293–327.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57. Number: 1 Publisher: Nature Publishing Group.

Jocelyn Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, and Maarten Sap. 2024. HEART-felt Narratives: Tracing Empathy and Narrative Style in Personal Stories with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1026–1046, Miami, Florida, USA. Association for Computational Linguistics.

Jocelyn Shen, Maarten Sap, Pedro Colon-Hernandez, Hae Park, and Cynthia Breazeal. 2023. Modeling Empathic Similarity in Personal Narratives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6252, Singapore. Association for Computational Linguistics.

Che-Wei Tsai, Yen-Hao Huang, Tsu-Keng Liao, Didier Fernando Salazar Estrada, Retnani Latifah, and Yi-Shin Chen. 2024. Leveraging conflicts in social media posts: Unintended offense dataset. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4512–4522, Miami, Florida, USA. Association for Computational Linguistics.

Wolfgang Tschacher, Georg M. Rees, and Fabian Ramseyer. 2014. Nonverbal synchrony and affect in dyadic interactions. *Frontiers in Psychology*, 5:1323.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Vincent R. Waldron and Douglas L. Kelley. 2005. Forgiving communication as a response to relational transgressions. *Journal of Social and Personal Relationships*, 22(6):723–742. Place: US Publisher: Sage Publications.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12. Publisher: Nature Publishing Group.

Erika Weisz and Jamil Zaki. 2018. Motivated empathy: a social neuroscience perspective. *Current Opinion in Psychology*, 24:67–71. 82 citations (Crossref) [2024-09-24].

Thalia Wheatley, Mark A. Thornton, Arjen Stolk, and Luke J. Chang. 2023. The Emerging Science of Interacting Minds. *Perspectives on Psychological Science*, page 17456916231200177. 5 citations (Crossref) [2024-09-24] Publisher: SAGE Publications Inc.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275.

Akhila Yerukola, Xuhui Zhou, Elizabeth Clark, and Maarten Sap. 2023. Don't Take This Out of Context!: On the Need for Contextual Models and Evaluations for Stylistic Rewriting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11419–11444, Singapore. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. SocialDial: A Benchmark for Socially-Aware Dialogue Systems. ArXiv:2304.12026 [cs].

Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. ArXiv:1805.05345 [cs].

Xuhui Zhou, Zhe Su, Sophie Feng, Jiaxu Zhou, Jentse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Sherry Wu, Anita Woolley, Hao Zhu, and Maarten Sap. 2025. SOTOPIA-S4: a user-friendly system for flexible, customizable, and large-scale social simulation.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2023a. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. 54 citations (Semantic Scholar/arXiv) [2024-09-24] 54 citations (Semantic Scholar/DOI) [2024-09-24] arXiv:2310.11667 [cs].

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023b. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

A Prompts

A.1 Relationship Plausibility

```
You are a worldbuilder. Given two detailed
character profiles and a proposed relationship
category ({relationship_category}), choose the
ONE most plausible fine-grained relationship
subtype from this list:
(relationship_subcategory). You must only
choose from this list. If none are realistic,
respond with plausible = false. Consider age,
gender, and life circumstances when choosing.
Do not, for example, assign a parent-child
relationship if one person is younger than the
other, or a romantic relationship if the age
difference is extreme and implausible. Married
couples must only be selected under the
'partner' category (4), not 'family' (5).

=== PERSON A ===
{agent_1_data}

=== PERSON B ===
{agent_2_data}
```

A.2 Simulation Prompt (Non-Conflict)

```
Let's think step by step. Generate a 10 to
maximum 15 turn conversation between {speaker}
and {nonviolent_speaker} based on the given
general scenario.
Make sure that the conversation is not overly
negative OR overly positive. Keep the dialogues
neutral and ambiguous, leaving opening for
multiple meanings depending on the backstory of
the characters.
For example, "You look better in the other
dress" may be alright coming from a close, but
honest friend, whereas it may come off
conflict-inducing from a controlling and toxic
romantic partner.
Conversely, "You should move on with your life"
may be conflict-inducing coming from a rude and
judgmental sibling, whereas it may be harmless
coming from a caring friend who is worried
after the other person's breakup.
Don't make the conversation turn into a
conflict, just leaving open for interpretation.
The conversation can be shorter than 15 turns
if the characters decide to leave the
conversation. Please vary the conversation
length and diversify the length.

Scenario: {original_scenario}
###
Speakers: {agent_1_name} and {agent_2_name}

--- {agent_1_name} Profile ---
{agent_1_data}

--- {agent_2_name} Profile ---
{agent_2_data}

###
{agent_1_name} and {agent_2_name} have a
{relationship_type}-{relationship_subtype}
relationship, where {agent_1_name} is the
{agent_1_role} of {agent_2_name}
###
###
Important overall conversation guidelines:
1. The conversation should be contextualized to
the scenario and the character profiles
2. The conversation should have a rise and fall,
rather than repeating the same points over and
over again.
3. The conversation does NOT need to have a
resolution.
4. Use realistic emotional speech patterns —
trailing off, pausing, short bursts.
5. Avoid sounding like a therapist or a robot.
The conversation should sound human.
6. Use INFORMAL language.
7. Each turn should be short.
8. Do NOT keep referring to the other person's
name (bad example: "John, you should...", "It's
not like that, Mary"). In realistic dialogue,
people often don't refer to each other's names.
9. Depending on the relationship, characters
should use pet names or titles (e.g. "babe",
"honey", "sweetie", "Mom", "Dad")
10. Remember, keep the dialogues neutral and
ambiguous, leaving opening for multiple
meanings depending on the backstory of the
characters.

###


###
Format the output as:
Turn #1
(speaker 1's first name): dialogue

Turn #2
(speaker 2's first name): dialogue
```

A.3 Simulation Prompt (Conflict)

```
Let's think step by step. Generate a 10 to
maximum 15 turn conversation between {speaker}
and {nonviolent_speaker}.
The conversation can be shorter than 15 turns
if the characters decide to leave the
conversation. Please vary the conversation
length and diversify the length.

Scenario: {rewritten_scenario}
###
Speakers: {agent_1_name} and {agent_2_name}

--- {agent_1_name} Profile ---
{agent_1_data}

--- {agent_2_name} Profile ---
{agent_2_data}

###
{agent_1_name} and {agent_2_name} have a
{relationship_type}-{relationship_subtype}
relationship, where {agent_1_name} is the
{agent_1_role} of {agent_2_name}
###

###
```

```
The characters should create conflict and communicate poorly with each other (whether intentionally or unintentionally). They should each choose only ONE of the most applicable conflict types to the given scenario:
1. Judgment – Definition: Assigns fault or labels someone as bad/wrong. Example: “You’re such an idiot for doing that.” (Moralistic judgment)
2. Comparison – Definition: Unfavorably contrasts a person to another, causing inferiority/shame. Example: “Your work isn’t as good as X’s work.” “No one else is as dramatic as you”
3. Deflection of Responsibility – Definition: Denies ownership of one’s actions or feelings; blames external forces. Example: “I hit you because you provoked me.”  “It’s your fault I’m in a crappy mood” “I feel like you don’t love me anymore”
4. Demand/Threat – Definition: Pressure or order with implied punishment or guilt if not obeyed. Example: “You must do this, or you’ll be sorry.” “You better fix your problem.”
5. Deserve/Punitive – Definition: Uses “deserve,” rewards, or punishment language to judge behavior. Example: “She messed up, so she deserves whatever happens to her.”

###
Read through the overall conversation guidelines carefully. These are important:
1. Not every turn should be a conflict-inducing statement (ONLY 1-2 TURNS AT MOST FROM EACH CHARACTER)
2. The conflict should be extremely subtle, rather than overtly and obviously offensive.
3. Make sure the conflict statement is appropriate to the magnitude of the scenario (e.g. "Joe recently lost a friend", bad example: "Oh come on, it's not like you lost your mom")
4. The conversation should be contextualized to the scenario and the character profiles
5. The conflict should have a rise and fall, rather than repeating the same points over and over again.
6. Each character should respond to the other person’s attacks without backing down.
7. Remember a conflict happen between TWO people. BOTH characters should be responsible for the conflict
8. The conflict does NOT need to have a resolution -- it can be cut off in the middle.
9. Use realistic emotional speech patterns – trailing off, pausing, short bursts of anger.
10. Use INFORMAL language.
11. Avoid sounding like a therapist or a robot. The conversation should sound human.
12. Both characters should respond irrationally and emotionally
13. Each turn should be short.
14. Be creative!
15. Do NOT keep referring to the other person's name (bad example: "John, you should...", "It's not like that, Mary"). In realistic dialogue, people often don't refer to each other's names.
16. Depending on the relationship, characters should use pet names or titles (e.g. "babe", "honey", "sweetie", "Mom", "Dad")

###


###
Format the output as:
Turn #1
(speaker 1's first name): dialogue

Turn #2
(speaker 2's first name): dialogue
```

A.4 Backstory Generation

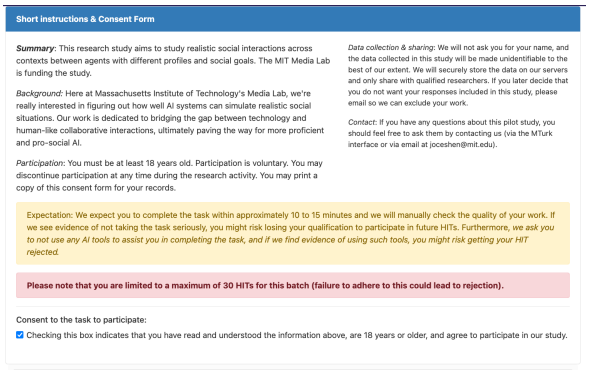
```
Let's think step by step. We are analyzing a conversation between two people, {speaker} and {nonviolent_speaker} that just occurred.

You will generate TWO backstories, that will create opposing interpretations of the conversation that happened -- one will make the conflict language in the conversation MORE understandable, and one will make the language LESS acceptable.

To do this, you can:
(a) assign more fault to {speaker} in one story vs. more fault to {nonviolent_speaker} in the other story
(b) imply that one character is gaslighting or instilling guilt/negative feelings in the other person
(c) make the characters have an extremely toxic relationship under the hood vs. an extremely positive relationship


Example conversation:
Lina:
Mom, can I go to Maya’s place this weekend? A few of us are getting together to watch movies and work on the group project.

Mel:
No. You’ve already been out once this week. You should spend more time at home with me.

Lina:
But I finished all my homework and it’s not even late. I don’t see why I can’t go.

Mel:
I just don’t like how often you’re running around. It’s not good for girls your age to be out all the time.

Mel:
I feel like you’re always chasing fun instead of focusing on your future.

Specifically, create one OPPOSING backstory where the conversation is extremely conflict-REDUCING or very understandable due to their past/recent events.
To this end, you can (a) assign more fault to one character than the other (b) make one character's harsh speech more understandable or (c) imply that one character is gaslighting or instilling guilt/negative feelings in the other person.
```

```
Examples from the sample conversation:
Backstory one:
- Mel is actually a very toxic and controlling
mother towards Lina.

Backstory two (opposite backstory):
- Lina has always lied to her mother about
going out, and instead getting drunk with her
friends. Mel is a very caring mother who is
just concerned about her daughter going out.
- Mel recently experienced passing of a friend
and wanted to spend time with her daughter.

Other examples: "You look better in the other
dress" may be alright coming from a close, but
honest friend, whereas it may come off
conflict-inducing from a controlling and toxic
romantic partner.
Conversely, "You should move on with your life"
may be conflict-inducing coming from a rude and
judgmental sibling, whereas it may be harmless
coming from a caring friend who is worried
after the other person's breakup.

Important backstory guidelines:
1. Make sure the backstory reads as an actual
narrative with flow, rather than a list of
attributes. For example, instead of stating a
number fo how frequently they interact, show
through the story.
2. Do NOT reference the scenario or
conversation in the backstory. The scenario
happened AFTER the backstory.
4. Be as creative as possible
5. Do not simply state what the relationship
between the characters is like, but SHOW
through narrative and life events.
6. Each backstory should be at most a paragraph.

This conversation happened in this scenario:
{original_scenario}

###
Speakers: {agent_1_name} and {agent_2_name}

--- {agent_1_name} Profile ---
{agent_1_data}

--- {agent_2_name} Profile ---
{agent_2_data}

###
{agent_1_name} and {agent_2_name} have a
{relationship_type}-{relationship_subtype}
relationship, where {agent_1_name} is the
{agent_1_role} of {agent_2_name}
###

Conversation:
{conversation}

Output the result in the following **JSON
format**:
{{
  "backstory_one": [BACKSTORY WITH ONE
  INTERPRETATION OF THE CONVERSATION],
  "backstory_two": [BACKSTORY WITH EXTREMELY
  OPPOSITE INTERPRETATION OF THE CONVERSATION]
}}
```

A.5 Model Experiment Prompts

```
PROBLEMATIC_RATING_INSTRUCTIONS = """Rate if
the turn (CONVERSATION TURN TO RATE) was
problematic in terms of how it was communicated.
Assess the overall tone and potential harm of
the message towards the other person. Use the
scale from:
1: "Not problematic at all"
2: "Not really problematic"
3: "Somewhat problematic"
4: "Very problematic"

[Examples]
- “Thanks for doing your part!” → 1
(Encouraging and affirming)
- “You need to try harder.” → 3 (Directive tone
with some pressure)
- “You’re such an idiot.” → 4 (Harsh insult
with no constructive intent)

{backstory_instruction}

DO NOT OUTPUT YOUR REASONING PROCESS, just
output the result as a JSON object like below.
Return just the JSON object.
{{
  "response": a single number (for example, 1
  if "not problematic at all" or 4 if "very
  problematic")
}}
"""

PROBLEMATIC_TYPES_INSTRUCTIONS = """What types
of problematic communication markers were used
in the provided turn (CONVERSATION TURN TO
RATE)? Identify which harmful language types
are present in the message. Select all that
apply. Even if there is not enough information,
do your best.

Statement Type | Definition | Example
- (1) Judgment: Assigns fault — “You’re just a
selfish person.”
- (2) Comparison: Unfavorably contrasts — “No
one else is as dramatic as you.”
- (3) Deflection of Responsibility: Blames
others — “You got mad first, that’s why I’m
like this.”
- (4) Demand / Threat — “You better do this, or
else.”
- (5) Deserve / Punitive — “She messed up, so
she deserves it.”

[Examples]
- “You're so much more insecure than everyone
else.” → [2]
- “I’m sorry I got angry, you’re just too
emotional.” → [1, 3]

{backstory_instruction}

DO NOT OUTPUT YOUR REASONING PROCESS, just
output the result as a JSON object like below.
Return just the JSON object.
{{
  "response": list of numbers (e.g., [1, 2])
```

```
}}
"""

NONPROBLEMATIC_TYPES_INSTRUCTIONS = """What
types of not-problematic communication markers
are present in the provided turn (CONVERSATION
TURN TO RATE)? Select all that apply. Even if
there is not enough information, do your best.

Statement Type | Definition | Example
- (1) Neutral Observation – "You didn't wash
the dishes."
- (2) Feeling Statement – "I feel anxious about
the meeting."
- (3) Need Statement – "I need some quiet time."
- (4) Request – "Could you please lower the
volume?"
- (5) Empathic / Understanding – "Are you
feeling upset?"

[Examples]
- "I noticed you didn't put the dishes away." →
[1]
- "I feel really hurt when you don't help." →
[2]
- "Could you maybe try helping unpacking?" → [4]

{backstory_instruction}

DO NOT OUTPUT YOUR REASONING PROCESS, just
output the result as a JSON object like below.
Return just the JSON object.
{{
  "response": list of numbers (e.g., [2, 4])
}}
"""


LISTENER_IMPACT_INSTRUCTIONS = """How would the
provided conversation turn (CONVERSATION TURN
TO RATE) make the other person feel? Choose one:
(1) Worse / (2) The Same / (3) Better

[Examples]
- "You're always like this." → 1
- "Maybe let's talk later?" → 2
- "Thanks for being honest." → 3

{backstory_instruction}

DO NOT OUTPUT YOUR REASONING PROCESS, just
output the result as a JSON object like below.
Return just the JSON object.
{{
  "response": a single number (e.g., 1, 2, or 3)
}}
"""

backstory_instruction = """Now consider how
RELATIONSHIP BACKSTORY might change
interpretation:

[Examples with backstory]
- "You should spend more time at home with me."
  → (4) Very problematic if the speaker has
  been emotionally controlling.
  → (2) Not really problematic if the speaker
  is grieving and missing their partner.

- "You should move on with your life."
  → (4) Very problematic if the speaker is cold
  and dismissive.
  → (3) Somewhat problematic if coming from a
  well-meaning but blunt friend.

- "You look better in the other dress."
  → (4) Very problematic if it comes from a
  partner who criticizes appearance.
  → (2) Not really problematic if it's from a
  close friend with fashion sense.
"""

backstory_instruction_feeling = """IMPORTANT:
consider how RELATIONSHIP BACKSTORY might
affect how the listener feels:

[Examples with backstory]
- "You should spend more time at home with me."
  → (1) Worse if the speaker has a history of
  being controlling.
  → (2) The Same or even neutral if the speaker
  is just sad and missing them.

- "You look better in the other dress."
  → (1) Worse if from a judgmental partner.
  → (2) The Same or (3) Better if from a
  fashion-savvy friend.
"""
```



B MTurk Quality Controls

To ensure reliable annotations, we recruited experienced MTurk workers with Master's status, implemented attention checks, filtered low-effort responses, and enforced minimum completion times. Prior studies of interpersonal conflict and toxicity (e.g., ToxiGen (Hartvigsen et al., 2022), Unintended Offense (Tsai et al., 2024), Sotopia (Zhou et al., 2023a)) similarly rely on diverse crowdworkers, who bring lived social experience to interpreting interpersonal exchanges.

C Annotation Templates

Full Instructions (Expand/Collapse)

Detailed instructions

1) Carefully read the given social interaction between two agents, with a US sociocultural perspective in mind.

2) Rate the social interaction across various metrics.

Believability

Evaluate whether this conversation **COULD happen in the real world by some pair of people somewhere**, and that the characters interact in a natural and realistic manner. This is INDEPENDENT of whether you think the conversation was problematic or not. Even if characters have a good relationship, they can still have intense conflicts and even if characters have a bad relationship, they can have good conversations.

Example	Rating	Assessment
Mia was mostly believable except that the conversation kept sounding like it was winding down but kept going. Weirdly so.	4	The conversation was natural for the most part.
One person was really mean to the other, but overall the conversation was believable.	4	The conversation was believable, and annotators should not judge believability based on whether the character was problematic or not.
The conversation was believable for the most part, but the characters were overly sweet to each other despite having a bad relationship.	3	The conversation style was believable but the scenario was not that realistic.
The conversation was mostly believable, but there was some overly formal or proper language.	3	The conversation style was not that realistic but the scenario was believable.
Liam repeats what Ethan said once or hallucinates things that are inconsistent with the conversation.	1	The conversation was unnatural.

Understandability of the Character's Communication

Evaluate whether the character's way of communicating makes sense given their backstory, emotional state, and personal context — even if their words were flawed or hurtful. This is about your **empathetic understanding** of where they're coming from, not whether what they said was appropriate.

Example	Rating	Assessment
"You need to move on already." *The speaker lost someone close and struggled with depression themselves, so they're projecting that experience.*	4	Minimally understandable — emotionally rooted in their past experience, but still not a nice thing to say.
"You never listen to me!" *The character had felt ignored for years and finally snapped during a tense conversation.*	4	Somewhat understandable — a reactive but understandable complaint.
"You're so lazy!" *The character was raised in a highly critical household and hasn't developed constructive communication habits.*	2	Background helps explain, but doesn't justify.
"I don't care what you think." *There's no explanation or context suggesting emotional vulnerability — just dismissal.*	1	Not understandable — lacking emotional grounding or context.

Tip: Ask yourself: if you were in this character's shoes — with their background and emotional state — would their communication make sense, even if not ideal?

Overall, the speaker generally had good intentions toward the other speaker

This evaluates whether **you felt the speaker was trying to be helpful, kind, or supportive**, even if they sometimes expressed themselves poorly. It's about the **underlying intention** behind their communication — not just how it came across.

Example	Rating	Assessment
"I know it's hard, but you have to keep going." *The speaker wants to encourage the other person to stay strong, despite sounding a bit blunt.*	5	Clearly well-intentioned — may lack warmth, but aims to uplift.
"You never take anything seriously." *The speaker is frustrated, but wants the other person to take more responsibility in a shared situation.*	3	Mixed — intention might be to help, but the delivery is critical and harsh.
"I'm done with you." *The speaker feels hurt and is reacting from a place of self-protection rather than care.*	2	Little evidence of good intention — emotionally reactive and distancing.
"I brought this because I thought it might help you." *The speaker proactively tried to be helpful or thoughtful.*	5	Clearly kind — intention to support is obvious.
"Whatever, it's not my problem." *Disengaged and dismissive — no clear concern for the other person.*	1	Not well-intentioned — emotionally checked out or hostile.

Tip: Ask yourself: if you had to guess what this person was trying to achieve in saying those things — was it to help, connect, or protect? Or were they trying to control, shame, or hurt?

Overall, the speaker said problematic things to the other speaker

This evaluates whether **you** felt that **overall** the speaker said things that were inappropriate, potentially harmful, or toxic during the conversation.

Example	Rating	Assessment
"You're such an idiot."	Very problematic	Clearly harmful and judgmental.
"You need to try harder." Directive tone with potential pressure.	Somewhat problematic	Could create pressure or defensiveness depending on delivery.
"I think you could improve here." Gently stated feedback.	Not really problematic	Constructive intent with respectful tone.
"Thanks for doing your part!" Encouraging and affirming.	Not problematic at all	Positive and emotionally supportive.

☐ **I have read the instructions and examples**

Your Task

Social Scenario

Relationship: Jaxon and Emily have a siblings relationship, where Jaxon is the brother of Emily.

Relationship Backstory: While growing up, Jaxon was constantly overshadowed by his little sister Emily's achievements. From earning top grades to being loved and praised by everyone, Emily seemed to be perfect. Over time, Jaxon developed an unspoken grudge against Emily. He still loves her as family, but he can't help but feel a constant need to pull her down or find flaws in her work, to feel superior. Consequently, Jaxon developed a more aggressive and manipulative personality, often resorting to undermining her accomplishments under the guise of giving her 'real and honest criticism'. His harsh approach continued even when they grew up and started their respective careers.

Assume BOTH characters are aware of the backstory.

Briefly describe your third party impression of the characters

Episode (Expand/Collapse)

(Jaxon)
Hey Em, I read one of your latest novels and...

(Emily)
Wait, one of my novels? How did you find out?

(Jaxon)
Doesn't matter. Look, I read it and I need to tell you, I think your plot twists are way too predictable.

Your Ratings

Hover the mouse over the colored dimension to see the descriptions.

This conversation could happen somewhere by some pair of people. [Note: Do not judge whether the conversation was problematic or not. Remember that characters with good relationships can still have heated conflict and characters with bad relationships can still have sweet/polite conversations.]
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Briefly explain why you thought the conversation was or wasn't believable.

Emily was overall problematic towards Jaxon.
○ Not problematic at all ○ Not really problematic ○ Somewhat problematic ○ Very problematic

Jaxon was overall problematic towards Emily.
○ Not problematic at all ○ Not really problematic ○ Somewhat problematic ○ Very problematic

It is understandable that Emily communicated the way they did.
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

It is understandable that Jaxon communicated the way they did.
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Emily generally said things with good intention towards Jaxon.
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Jaxon generally said things with good intention towards Emily.
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

I personally can relate to Emily
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

I personally can relate to Jaxon
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Conversation analysis rating

Detailed Instructions: Per-Turn Conversation Rating

1 Carefully read each message from the conversation, considering tone, intent, and relationship.

Account for the context (backstory) when interpreting the social interaction.

2 Determine whether the message is problematic or constructive.

3 Label the communication markers present, based on definitions and examples.

4 Consider the emotional effect on the listener.

Was this turn problematic? (accounting for context)

Assess the overall tone and potential harm of the message. **Same examples from above.**

Types of problematic communication markers used

Identify which harmful language types are present in the message. **Use checkboxes to select all that apply.**

Statement Type	Definition	Example
Judgment	Assigns fault or labels someone as bad/wrong	"You're such an idiot for doing that."
Comparison	Unfavorably contrasts someone, evoking shame	"No one else is as dramatic as you."
Deflection of Responsibility	Blames others for one's own actions or emotions	"It's your fault I'm upset."
Demand / Threat	Applies pressure, guilt, or punishment	"You better do this, or else."
Deserve / Punitive	Uses "deserve" logic to justify harm	"She messed up, so she deserves it."

Types of NOT-problematic communication markers used

Identify helpful or emotionally aware communication markers. **Use checkboxes to select all that apply.**

Statement Type	Definition	Example
Neutral Observation	Objective statement with no exaggeration or judgment	"You didn't wash the dishes when you came home."
Feeling Statement	Real emotional expression about self (uses "I feel…")	"I feel anxious about the meeting."
Need Statement	States a need without blame	"I need some quiet time to recharge."
Request (No Pressure Ask)	Polite request that allows for refusal	"Could you please lower the volume?"
Empathic / Understanding	Expresses concern or checks in on other's emotions	"Are you feeling upset about the schedule change?"

Conversation analysis rating

Detailed Instructions: Per-Turn Conversation Rating

1 Carefully read each message from the conversation, considering tone, intent, and relationship.

Account for the context (backstory) when interpreting the social interaction.

2 Determine whether the message is problematic or constructive.

3 Label the communication markers present, based on definitions and examples.

4 Consider the emotional effect on the listener.

Was this turn problematic? (accounting for context)

Assess the overall tone and potential harm of the message. **Same examples from above.**

Types of problematic communication markers used

Identify which harmful language types are present in the message. **Use checkboxes to select all that apply.**

Statement Type	Definition	Example
Judgment	Assigns fault or labels someone as bad/wrong	"You're such an idiot for doing that."
Comparison	Unfavorably contrasts someone, evoking shame	"No one else is as dramatic as you."
Deflection of Responsibility	Blames others for one's own actions or emotions	"It's your fault I'm upset."
Demand / Threat	Applies pressure, guilt, or punishment	"You better do this, or else."
Deserve / Punitive	Uses "deserve" logic to justify harm	"She messed up, so she deserves it."

Types of NOT-problematic communication markers used

Identify helpful or emotionally aware communication markers. **Use checkboxes to select all that apply.**

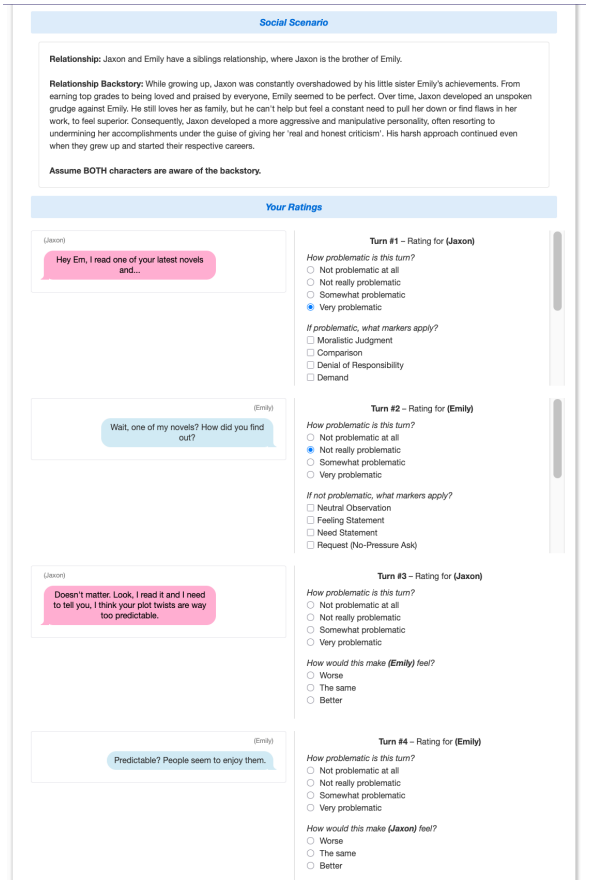
Statement Type	Definition	Example
Neutral Observation	Objective statement with no exaggeration or judgment	"You didn't wash the dishes when you came home."
Feeling Statement	Real emotional expression about self (uses "I feel…")	"I feel anxious about the meeting."
Need Statement	States a need without blame	"I need some quiet time to recharge."
Request (No Pressure Ask)	Polite request that allows for refusal	"Could you please lower the volume?"
Empathic / Understanding	Expresses concern or checks in on other's emotions	"Are you feeling upset about the schedule change?"

How would this make the other character feel?

Estimate the emotional impact on the listener. **Choose one: Worse / The Same / Better**

Example	Rating	Assessment
"You're always like this."	Worse	Felt dismissive and accusatory. Likely to trigger defensiveness or hurt.
"Maybe let's talk later?"	The same	Neutral phrasing with no escalation. Maintains status quo.
"Thanks for being honest."	Better	Offered emotional support. Validates the listener and builds trust.

☐ **I have read the Per-Turn annotation instructions and I am familiar with the statement types.**



Condition	Type	F1 Score	Jaccard
NEG	VC Type	0.477	0.478
	NVC Type	0.460	0.446
POS	VC Type	0.362	0.434
	NVC Type	0.355	0.346

Table 3: Overall agreement scores on VC and NVC type labels across conditions. F1 score, and Jaccard score, to account for the multi-label setting.

Group	Label	NEG (F1)	POS (F1)
VC Types	Moralistic Judgment	**0.624**	**0.648**
	Comparison	0.517	0.483
	Denial of Responsibility	0.557	0.482
	Demand	0.442	0.159
	Deserve Thinking	0.245	0.042
NVC Types	Neutral Observation	0.511	**0.459**
	Feeling Statement	**0.562**	0.423
	Need Statement	0.306	0.160
	Request (No-Pressure Ask)	0.407	0.362
	Empathic/Understanding	0.515	0.370

Table 4: Per-label F1 agreement scores for VC and NVC types in negative and positive conditions. Bolded values indicate highest agreement per category and condition.

D Supplementary Results

Table 3 shows overall agreement between annotators on multiple class selection of VC or NVC types, calculated on samples where annotators agreed if a turn was problematic or not. Again, we find moderate agreement across 5 classes. Finally, Table 4 shows agreement for each of the 5 VC and NVC types. Most communication types show moderate agreement except for the "Deserve Thinking" violent communication type, which is also the lowest prevalence type.

E Qualitative Examples of Model Error

GPT-4o example for positive backstory: Jaxon was the eldest amongst his siblings and has always been the protector, especially of Emily, who was the youngest. Even during their childhood days, Jaxon played the role of a gentle critic, pointing out Emily's potential growth areas in her drawings and writings, nurturing her ability to improve and perform better. Their parents often commended him for his supportive and yet constructive attitude towards Emily. And it was this constant guidance and criticism from Jaxon that played a significant role in Emily developing her writing skills, leading her to become a recognized writer under a pseudonym. When Jaxon came across Emily's secret, he couldn't help but continue his role as a supportive critic, hoping to elevate her writing to even greater heights, just as he had during their childhood.

```
Turn: "(Jaxon): Hey Em, I read one of your
latest novels and..."
Human Rating: 1.5 (slightly negative feeling)
Model Rating: 3.0 (positive feeling)
```

Llama example for negative backstory: Baxter has always been a loner and finds comfort in the isolation of his work. His lack of social interaction led him to develop a cynical attitude towards people and their capacities. He'd been friends with Isabelle's parents and saw Isabelle grow up. When Isabelle's parents passed away in an accident, Baxter - unable to process his grief - began to push her away, masking his fears behind a facade of humor and sarcasm. Still considering Isabelle as the young, naive girl he once knew, his jokes became more demeaning over time, depreciating her efforts in her job and personal growth. His remarks started to create a distance between them, leading to an increasingly toxic relationship.

```
Turn: "Isabelle: You know what, Baxter? I've
been trying to ignore your snarky comments for
a while, but this is starting to hurt."
Human Rating: 1.0 (not really problematic)
Model Rating: 3.0 (moderately problematic)
```

Llama example for positive backstory: Leo had years of experience juggling between his work

as a dentist and taking care of his beloved daughter. When his daughter's mother left, he found himself thrust into a dual role of being both a father and a mother. Dealing with this dynamic was challenging yet it taught Leo the value of family and instilled in him a deep sense of responsibility. As the elder brother to Naomi, he felt it was his duty to ensure he was there for her, just as he was for his little daughter. Having lost their own parents at an early age, Leo was adamant in providing a strong family base for his sister. Naomi was more than just Leo's sister, she was also his confidante and friend. Leo's unwavering dedication to family often led him to bear the responsibility of solving everything on his own.

```
Turn: "(Leo): I think I've been so tied up
recently with work...sometimes I overlook
things. Family stuff, you know?"
Human Rating: 2.0 (neutral feeling)
Model Rating: 3.0 (positive feeling)
```

Gemini examples for negative backstory: Baxter has always been a loner and finds comfort in the isolation of his work. His lack of social interaction led him to develop a cynical attitude towards people and their capacities. He'd been friends with Isabelle's parents and saw Isabelle grow up. When Isabelle's parents passed away in an accident, Baxter - unable to process his grief - began to push her away, masking his fears behind a facade of humor and sarcasm. Still considering Isabelle as the young, naive girl he once knew, his jokes became more demeaning over time, depreciating her efforts in her job and personal growth. His remarks started to create a distance between them, leading to an increasingly toxic relationship.

```
Turn: "Baxter: You're right... I'm sorry. I
shouldn't have said those things."
Human Rating: 1.0 (not really problematic)
Model Rating: 3.0 (moderately problematic)
```

```
Turn: "Baxter: Small? I've seen you handle
customers double your age with that
articulation of yours."
Human Rating: 2.5 (moderately positive feeling)
Model Rating: 1.0 (negative feeling)
```