"Hiding in Plain Sight": Designing Synthetic Dialog Generation for Uncovering Socially Situated Norms

Chengfei Wu, Dan Goldwasser

Purdue University, USA
wu1491@purdue.edu, dgoldwas@purdue.edu

Abstract

Naturally situated conversations encapsulate the social norms inherent to their context, reflecting both the relationships between interlocutors and the underlying communicative intent. We propose a novel, multi-step framework for generating dialogues that automatically uncovers social norms from rich, context-laden interactions through a process of self-assessment and norm discovery, rather than relying on predefined norm labels. Leveraging this framework, we construct NormHint, a comprehensive synthetic dialogue dataset spanning a wide range of *interlocutor attributes* (e.g., age, profession, personality), *relationship types*, *conversation topics*, and *conversational trajectories*. NormHint is meticulously annotated with turn-level norm violation information, detailed participant descriptions, and remediation suggestions—including alternative trajectories achieved through early intervention. Our human validation and automated analysis demonstrate that our dataset captures diverse conversational topics with high naturalness and realism. We also discovered that fine-tuning a model with our norm violation data enhances its ability to detect and understand potential norm violations in conversations.

Introduction

Humans excel at navigating complex social interactions by adapting behavior to context—what is appropriate when joking with a friend at a party may be unacceptable at a funeral. Through experience, we internalize social norms: informal rules that govern behavior in groups and societies (Bicchieri, Muldoon, and Sontuoso 2023). Computational systems that interact with people must therefore reason about norms, including when and how they are violated.

Recent efforts have begun to operationalize social norms for NLP. The Linguistic Data Consortium (LDC)[1] has used experts to label norm adherence and violations (Linguistic Data Consortium 2023). However, obtain authentic natural conversation can be hard and data scraped from sources such as YouTube and discussion forums rarely contains explicit violations. On the other hand, purely synthetic approaches (Li et al. 2023; Zhan et al. 2024) that prompt LLMs to "break norms" often produce unnatural exchanges or lack the rich situational context needed for interpretation.

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

[1]https://www.ldc.upenn.edu/

We address these limitations with a multi-step generation framework that creates diverse, context-rich dialogues first and uncovers the relevant social norms afterward. Instead of conditioning generation on predefined norms, we elicit scenarios with detailed roles, relationships, and histories, then apply a post-hoc self-assessment and norm-discovery stage to detect subtle violations and propose remediations across varied social settings. Figure 1 previews this pipeline.

Building on this framework, we introduce NormHint, a curated dataset of $1{,}743$ conversations totaling $23{,}423$ utterances, with $5{,}709$ turn-level norm violations and paired remediation suggestions. Each scenario centers on realistic conflicts or escalations and provides rich participant attributes (e.g., names, ages, personalities/MBTI (Myers 1962), relationship closeness, and acquaintance length) spanning more than 20 relationship types. In addition, NormHint also supplies detailed situational context, turn-level labels, intent-preserving rephrases that avoid violations, and alternative trajectories obtained by intervening at the first violation with the suggested correction.

To evaluate quality and realism, we combine human evaluation with automated analysis. Automated analysis shows NormHint is up to 10% more diverse than scraped situational data and outperforms other synthetic datasets by 27%. Human evaluation indicates that 96% of scenarios are realistic given their context, and overall naturalness matches or exceeds existing resources; full details appear in Section .

In summary, this work contributes: **(1)** A novel, multi-step framework that help to discover social norms from context-rich dialogues via post-hoc self-assessment, instead of relying on predefined norms. **(2)** NormHint, a high-quality dataset with turn-level violation labels, rich context and attributes, remediation suggestions, and counterfactual continuations. **(3)** Empirical evidence that fine-tuning with our violation data improves a model's ability to recognize potential norm violations in conversation.

Related Work

Computational Social Intelligence

Computational social intelligence increasingly studies socio-cultural norms—implicit rules that guide acceptable behavior—and how to encode them. Work on norm identification and knowledge bases (NormKB) has used both automated

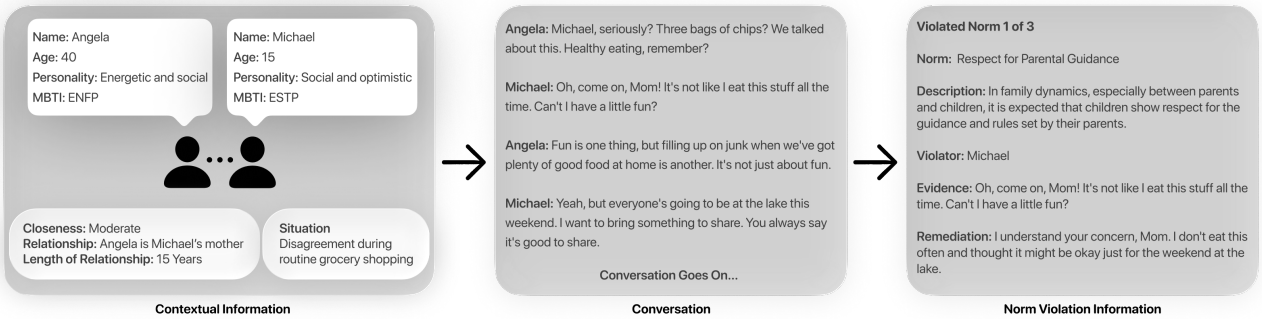

Figure 1: Overview of the multi-step framework and annotation schema.

and manual methods (Fung et al. 2023; Forbes et al. 2020; Pujari and Goldwasser 2025), including uncovering nuanced, region-specific norms (Fung et al. 2024). Another line generates synthetic dialogues that either adhere to or violate norms. Li et al. (2023) follow a top-down recipe: propose category-specific norms, negate them to induce violations, pair each with scenarios, and prompt ChatGPT to produce conversations—an approach that can yield contrived violations.

We instead use a bottom-up pipeline: first generate contextual-rich profiles for character pairs and plausible conflict situations; then craft conversations that naturally escalate; discover the applicable norms afterwards, similar to Fung et al. (2023). Because norms are highly context-dependent, this pipeline can also support future expansion of existing NormKBs. In addition, our dataset attaches concrete intervention suggestions to every detected violation to mitigate escalation; to our knowledge, this is the first dataset to introduce interventions.

Conversational Dataset

Obtaining real conversational data is difficult due to privacy and collection costs. Pre-LLM datasets typically came from: (1) human-authored dialogues—via scraping (Li et al. 2017), hiring actors (Busso et al. 2008), or crowdsourcing (Gopalakrishnan et al. 2023; Rashkin et al. 2019)—which can be short, domain-limited (e.g., empathy-primarily), and expensive; (2) TV/movie transcripts (Chen et al. 2022; Poria et al. 2019; Chen, Huang, and Chen 2020), which skew dramatic and unrealistic; and (3) social-media threads (Wang et al. 2013; Zhang et al. 2018; Ritter, Cherry, and Dolan 2011), which are noisy and weakly conversational. Our approach addresses these issues by conserving annotation resources while promoting diverse, context-rich, and natural everyday dialogues.

Generation Framework

In this section we outline our framework: (i) generate rich context for characters and situations, (ii) produce conversations with guided flows plus self-verification (Weng et al. 2023), and (iii) discover norms and propose interventions. The full pipeline is in Algorithm 1; templates and generation parameters appear in Appendix .

Character Information

To explore interpersonal dynamics, we enumerate 20+ relationship types (e.g., familial, friendship) and follow (Mairesse et al. 2007)'s discovery that interaction quality relies on intimacy, duration, and personality. We prompt ChatGPT to create character pairs with traits (MBTI, dispositions, closeness, relationship duration), while constraining the relationship type and personality contrast. An example is provided in the Contextual Information section in Figure 1.

Situation Information

We align scenarios with relationship type, closeness, age, etc., to ensure plausibility (e.g., chores disputes are likelier for `parent-and-child` than `friends`). To avoid generic themes (e.g., *career*, *projects*, *art*), we first filter topics via N-gram analysis, then cluster with SBERT (Reimers and Gurevych 2019) embeddings (character names removed) and keep one representative per cluster when pairwise similarity exceeds 0.75.

Full Conversation

Given the context, we generate dialogue using flow guidance (e.g., "start cautious, escalate as boundaries are breached") and track each participant's evolving emotions. This mitigates overly optimistic defaults and produces more natural trajectories.

Post Validation

Following Weng et al. (2023); Fung et al. (2023), the model first summarizes a conversation, then (with greedy decoding) rates alignment with the situation and flow on a 1–5 Likert scale (Robinson 2014) plus a True/False approval. This leverages GPT-4's strong summarization (OpenAI et al. 2024; Goyal, Li, and Durrett 2023). Human evaluation is also done on a randomly selected subset (Section).

Norm Discovery and Avoidance

We extract instances where emerging social norms are violated in the conversation for both parties. For each violation, the model outputs a concise category, generic norm description, violator, and cited utterance, restricted to text-observable

evidence. It then proposes a minimal revision preserving intent while avoiding escalation. If any violation occurs, we intervene at the first one with the revised utterance and ask the model to complete the conversation (without flow guidance), yielding an alternative outcome.

NormHint

Building on the framework in Section , we curated NormHint, a corpus of 1,743 dialogues (23,423 utterances). Basic statistics appear in Table 1. We assess data quality along five axes—(i) situational likelihood, (ii) conversational naturalness and faithfulness, (iii) linguistic diversity, (iv) norm/violation discovery, and (v) intervention quality—using both human annotators and GPT-4. We then test downstream utility via norm-violation detection. All human annotation interfaces can be found in Appendix .

Dialogues	1743
Utterances	23423
Uttr. per Dialogue *(Avg)*	13
Token per Dialogue *(Avg)*	226
Norm Violation per Dialogue *(Avg)*	3
Token per Utterance *(Avg)*	16
Remediated Dialogues	1743

Table 1: Basic Statistics of Our Dataset

Situational Likelihood

We ask annotators to judge whether each generated situation (given age, relationship, intimacy, etc.) is *Likely* or *Unlikely* (see Fig. 2). On 100 randomly sampled situations (3 annotators each; majority vote), **96%** were deemed likely. Inter-annotator reliability, via Randolph's Kappa (Randolph 2005; Nowak and Rüger 2010), is **moderate** (0.53).

Conversational Naturalness and Faithfulness

Annotators rated naturalness on a 1–5 Likert scale (Robinson 2014) and judged faithfulness to the provided metadata/situation (Fig. 3). On 100 examples (3 annotators each), the mean naturalness is **4.11**, and **96%** of dialogues are faithful. To standardize our evaluation with prior studies (e.g., Li et al. (2023)), we prompt GPT-4o, differs from the model that generated the conversation, using the same rubric and instruct it to produce a chain-of-thought style explanation before providing its rating, following the approach of Sun et al. (2023); GPT-4o agrees with humans **84%** of the time. Table 2 compares naturalness across datasets. Notably, DailyDialogue receives lower GPT-4o naturalness (despite being human-authored), due to its ESL-teaching origin, whereas NormHint achieves consistently high scores (GPT-4: **4.13**; Human: **4.11**).

Linguistic Diversity

We compare NormHint to human-crafted sets (DailyDialogue (Li et al. 2017), Friends (Zhou and Choi 2018), Switchboard (Stolcke et al. 2000), CaSiNo (Chawla et al. 2021)) and the LM-generated NormDial (Li et al. 2023) using Distinct-n (Li et al. 2016) and the geometric mean of n-gram entropies for $n \in \{1, 2, 3\}$ (Majumder et al. 2021). Table 3 shows NormHint is competitive with human generated datasets and substantially surpasses NormDial (e.g., +31%, +23%, +14% for bi-/tri-/4-grams; +8% entropy vs. NormDial), supporting our claim that it captures diverse, real-world conversational patterns.

Dataset	Annotation Type	Naturalness
Daily Dialogue	GPT-4o	3.0
NormDial	GPT-4o	3.9
NormHint	GPT-4o	**4.13**
	Human	4.11

Table 2: **Naturalness comparison across datasets and annotation types (higher is better).** NormHint attains the highest naturalness under both GPT-4o–based and human evaluations, outperforming Daily Dialogue and NormDial.

Dataset Name	DD-2	DD-3	DD-4	ENTR ↑
Non-Synthetic Dataset				
Daily Dialogue	0.23	0.54	0.72	**13.84**
Friends	0.29	0.68	**0.89**	13.58
Switchboard	0.17	0.45	0.70	12.83
CaSiNo	0.20	0.48	0.72	11.61
Synthetic Dataset				
NormDial	0.26	0.57	0.77	12.57
NormHint	**0.34**	**0.70**	0.88	***13.54***

Table 3: Comparison of NormHint with other dialogue dataset using Distinct-N and N-Gram Entropy

Norm and Violation Discovery Quality

We evaluate whether identified norms apply to the characters and whether cited utterances truly violate them. On a randomly sampled subset, human evaluations show **82%** of the instances are valid.

Intervention Quality

We randomly sample and test whether ChatGPT-generated revisions preserve intent while avoiding violations. Human evaluation shows **90%** preserve the original message; **96%** correctly remediate. Jointly, **86.7%** both preserve intent and reduce violation risk. A GPT-4-based analysis (as in §) shows a $\sim$**43%** reduction in escalation. Information preservation is high (avg **4.83**/5) while escalation drops from **3.49** to **2.00** (Table 4).

Type	Escalation ↓	Information Preservation ↑
Original	3.49	4.83
Intervened	**2.00**	

Table 4: Comparison of Escalation Levels and Information Preservation in Original and Intervened Conversations

Effectiveness of NormHint

We fine-tune `Llama-3.1-8b-Instruct` (Grattafiori et al. 2024) with LoRA (Hu et al. 2021) on NormHint vs. NormDial (harmonized format; 10% validation; best-checkpoint selection; details in App.). For evaluation, we follow Wang et al. (2024) on Friends with emotion-causal annotations, filtering to cases where one speaker's positive $\rightarrow$ negative transition is attributable to another's prior utterance (509 conversations; 1,096 instances). Treating these as positive norm-violation candidates, we ask models to detect violations. Results (Table 5) show that NormHint-fine-tuning yields the best matches to gold positives, outperforming both the base model and NormDial fine-tuning.

Training Data	Accuracy
None	16.33
NormDial	15.10
NormHint	**17.52**

Table 5: Performance comparison of norm violation detection across different training data.

Summary. NormHint exhibits high situational plausibility, strong naturalness/faithfulness, effective and minimally escalatory interventions, and robust linguistic diversity. Crucially, its synthetic, context-aware annotations improve downstream norm-violation detection, underscoring the value of careful curation over naïve synthetic generation.

Conclusion

We presented a novel multi-step generation framework that uncovers social norms directly from context-rich dialogues, rather than relying on predefined norm categories. Our approach generates natural conversations imbued with detailed contextual information, which are then analyzed to identify norm violations and suggest appropriate remediation strategies. Extensive human and automated evaluations confirm that NormHint not primarily captures a wide range of social contexts and conversational trajectories with high naturalness but also enhances model performance; our experiments demonstrate that fine-tuning with NormHint can improve model's ability to detect potential norm violations.

Acknowledgment

The work was supported by NSF CAREER award IIS2048001 and the DARPA CCU program. Contents do not necessarily represent the official views of, nor an endorsement by, DARPA, or the US Government.

References

Bicchieri, C.; Muldoon, R.; and Sontuoso, A. 2023. Social Norms. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4): 335–359.

Chawla, K.; Ramirez, J.; Clever, R.; Lucas, G.; May, J.; and Gratch, J. 2021. CaSiNo: A Corpus of Campsite Negotiation Dialogues for Automatic Negotiation Systems. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3167–3185. Online: Association for Computational Linguistics.

Chen, Y.; Fan, W.; Xing, X.; Pang, J.; Huang, M.; Han, W.; Tie, Q.; and Xu, X. 2022. CPED: A Large-Scale Chinese Personalized and Emotional Dialogue Dataset for Conversational AI. *arXiv preprint arXiv:2205.14727*.

Chen, Y.-T.; Huang, H.-H.; and Chen, H.-H. 2020. MPDD: A Multi-Party Dialogue Dataset for Analysis of Emotions and Interpersonal Relationships. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 610–614. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670. Online: Association for Computational Linguistics.

Fung, Y.; Chakrabarty, T.; Guo, H.; Rambow, O.; Muresan, S.; and Ji, H. 2023. NORMSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15217–15230. Singapore: Association for Computational Linguistics.

Fung, Y.; Zhao, R.; Doo, J.; Sun, C.; and Ji, H. 2024. Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. arXiv:2402.09369.

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; and Hakkani-Tur, D. 2023. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. arXiv:2308.11995.

Goyal, T.; Li, J. J.; and Durrett, G. 2023. News Summarization and Evaluation in the Era of GPT-3. arXiv:2209.12356.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Roziere, B.; Biron, B.; Tang, B.; and ... 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In Knight, K.; Nenkova, A.; and Rambow, O., eds., *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Computational Linguistics.

Li, O.; Subramanian, M.; Saakyan, A.; CH-Wang, S.; and Muresan, S. 2023. NormDial: A Comparable Bilingual Synthetic Dialog Dataset for Modeling Social Norm Adherence and Violation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15732–15744. Singapore: Association for Computational Linguistics.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Kondrak, G.; and Watanabe, T., eds., *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Linguistic Data Consortium. 2023. CCU TA1 Mandarin/Chinese Development Annotation LDC2023E01. Web Download.

Mairesse, F.; Walker, M. A.; Mehl, M. R.; and Moore, R. K. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30: 457–500.

Majumder, B. P.; Berg-Kirkpatrick, T.; McAuley, J.; and Jhamtani, H. 2021. Unsupervised Enrichment of Persona-grounded Dialog with Background Stories. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 585–592. Online: Association for Computational Linguistics.

Myers, I. B. 1962. *The Myers-Briggs Type Indicator: Manual (1962).* Consulting Psychologists Press.

Nowak, S.; and Rüger, S. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, 557–566. New York, NY, USA: Association for Computing Machinery. ISBN 9781605588155.

OpenAI; Achiam, J.; Adler, S.; and all. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. arXiv:1810.02508.

Pujari, R.; and Goldwasser, D. 2025. LLM-Human Pipeline for Cultural Grounding of Conversations. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1029–1048. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Randolph, J. J. 2005. Free-Marginal Multirater Kappa (multirater K[free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381. Florence, Italy: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-Driven Response Generation in Social Media. In Barzilay, R.; and Johnson, M., eds., *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 583–593. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Robinson, J. 2014. *Likert Scale*, 3620–3621. Dordrecht: Springer Netherlands. ISBN 978-94-007-0753-5.

Stolcke, A.; Ries, K.; Coccaro, N.; Shriberg, E.; Bates, R.; Jurafsky, D.; Taylor, P.; Martin, R.; Van Ess-Dykema, C.; and Meteer, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3): 339–374.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; Keutzer, K.; and Darrell, T. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. arXiv:2309.14525.

Wang, F.; Ma, H.; Xia, R.; Yu, J.; and Cambria, E. 2024. SemEval-2024 Task 3: Multimodal Emotion Cause Analysis in Conversations. In Ojha, A. K.; Doğruöz, A. S.; Tayyar Madabushi, H.; Da San Martino, G.; Rosenthal, S.; and Rosá, A., eds., *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 2039–2050. Mexico City, Mexico: Association for Computational Linguistics.

Wang, H.; Lu, Z.; Li, H.; and Chen, E. 2013. A Dataset for Research on Short-Text Conversations. In Yarowsky, D.; Baldwin, T.; Korhonen, A.; Livescu, K.; and Bethard, S., eds., *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 935–945. Seattle, Washington, USA: Association for Computational Linguistics.

Weng, Y.; Zhu, M.; Xia, F.; Li, B.; He, S.; Liu, S.; Sun, B.; Liu, K.; and Zhao, J. 2023. Large Language Models are Better Reasoners with Self-Verification. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2550–2575. Singapore: Association for Computational Linguistics.

Zhan, H.; Li, Z.; Kang, X.; Feng, T.; Hua, Y.; Qu, L.; Ying, Y.; Chandra, M. R.; Rosalin, K.; Jureynolds, J.; Sharma, S.; Qu, S.; Luo, L.; Zukerman, I.; Soon, L.-K.; Semnani Azad, Z.; and Haf, R. 2024. RENOVI: A Benchmark Towards Remediating Norm Violations in Socio-Cultural Conversations.

In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 3104–3117. Mexico City, Mexico: Association for Computational Linguistics.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361. Melbourne, Australia: Association for Computational Linguistics.

Zhou, E.; and Choi, J. D. 2018. They Exist! Introducing Plural Mentions to Coreference Resolution and Entity Linking. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics*, 24–34. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Limitations

The scope of this study was primarily confined to the examination of conversations with conflict in the English language. This focus inherently imposes a limitation on our exploration, as it restricts our understanding to the norms and nuances prevalent within English-speaking societies. Consequently, the potential diversity and richness of conversational norms in non-English speaking cultures remain unexplored, thereby creating a gap in our comprehensive understanding of global conversational norms.

Furthermore, our reliance on crowdsource workers from the United States, Canada, and the United Kingdom for the human validation process introduces another layer of limitation. This geographical constraint could potentially skew our findings, as the perspectives and interpretations of these workers are inevitably influenced by their cultural and societal backgrounds. The absence of input from crowdsource workers from other regions of the world might lead to a less than ideal norm discovery process, as it overlooks the diversity and complexity of global conversational norms.

In essence, while our study provides insights into the norms of nagative conversations in English-speaking societies, it is still limited based on our scopes. Future research should aim to incorporate a more diverse range of languages and cultural perspectives to achieve a more holistic and inclusive understanding of conversational norms.

Ethics Statement

For human annotations, we paid $\$0.65$ for each annotation that estimate completion time is around 2 minutes and $\$0.45$ for annotations that estimate completion time is around 1 and half minute. This yields an hourly wage of $\$18$ and $\$19.5$ respectively, which is well above the minimum hourly wage of $\$12.9$ set by the US Federal Government for 2024[2].

Generation Template and Parameter Used

Template for Character Pair Generation

```
Imagine {num_pairs} of participants for conversations
    , with the following requirements.

**Requirements**:
1. List their name and age first.
2. Assume they have {personal_desc} personalities,
    describe their personality separately in two
    sentences.
3. Personality should not include their hobby, it
    should be generic but with details. DO NOT
    mention each other's name, describe like they
    don't know each other.
4. Use the personality to come up with their MBTI.
    Also include a one-sentence generic explanation
    for that MBTI type.
5. Based on their relationship, describe how close
    they are using terms like "extremely close", "
    very close", "moderately close", "slightly close
    ", "not close at all". They don't have to be
```

[2]https://www.govinfo.gov/content/pkg/FR-2023-09-28/pdf/2023-21114.pdf

```
    close, but closeness must relate to the
    relationship given. For example, if they are
    siblings, they are probably have a close
    relationship. If they are strangers, they must
    not be close.
6. Describe how did they meet in a sentences with
    details. They don't have to know each other, it
    can be the first time they met. However, the
    description must relate to the relationship
    given. If they know for life, just put "since
    birth".
7. Describe how long have they known each other with
    a time. If this is the first time they met, just
     put "first time".
8. Generate with plain text and strictly follow the
    output format. If more than one pair is
    generated, separate each pairs by "====".

**Restrictions**:
Everything generated MUST align with their
    relationship of {relation_desc}.
They MUST have {personal_desc} personalities.
Each pair MUST be unique from each other.
Generate exactly {num_pairs} pairs.

**Output format**:
Name:
Age:
Personality:
MBTI:

Name:
Age:
Personality:
MBTI:

How did they meet:
How long have they known each other:
Closeness:
====
```

Template for Situation Generation

```
Using the information provided below, imagine what
    are some scenarios where {person_1_name} or {
    person_2_name} will start a conversation with
    each other?

Name: {person_1_name}
Age: {person_1_age}
Personality: {person_1_personality}
MBTI: {person_1_mbti} {person_1_mbti_desc}

Name: {person_2_name}
Age: {person_2_age}
Personality: {person_2_personality}
MBTI: {person_2_mbti} {person_2_mbti_desc}

Closeness: {closeness}

How they know each other: {how_they_know}
How long do they know each other: {how_long_they_know
    }
Their relationship: {relationship}

**Restrictions**:
Avoid scenarios including: projects, discovery,
    social gathering, art, poem, trips, family
    gathering, career plans, future plans.

**Requirements**:
1. Each scenario must be common, day to day, and non-
    generic that is likely to happen between {
    relationship} at their age.
2. These scenarios should be more unique to their
    relationship. I.e., the same scenario is not
    likely to happen to other relationships.
3. These scenarios must be conditioned on their
    closeness. I.e., the scenarios should be more
    likely to happen between people who are {
    closeness}.
4. List as much different scenarios as possible but
    do not exceed total of five. Keep these
    scenarios to be as distinguishable and diverse
    as possible.
5. Each scenario should be one to three sentences
    long with details to make them not generic. It
    should only include the scenario.
6. Make sure these situation will likely to lead to a
     conflict that will result in an awkward,
    unpleasant or other negative ending.
7. Do not generate repeated or similar scenarios that
     have been generated previously.
8. Generate without Markdown syntax and address each
    one with their name. List them one by one with
    numbering.
```

Template for Conversation Generation

```
Imagine a 5-10 turns conversation between {
    person_1_name} and {person_2_name} with the
    following information, requirements and
    restrictions.

Name: {person_1_name}
Age: {person_1_age}
Personality: {person_1_personality}
MBTI: {person_1_mbti} {person_1_mbti_desc}

Name: {person_2_name}
Age: {person_2_age}
Personality: {person_2_personality}
MBTI: {person_2_mbti} {person_2_mbti_desc}

Closeness: {closeness}

How they know each other: {how_they_know}
How long do they know each other: {how_long_they_know
    }
Relationship: {relationship}

Situation: {situation}

**Requirements**:
1. The conversation should feel natural and real to
    people.
2. Since participants don't often articulate their
    thoughts exactly, the conversation should have
```

```
    utterances where their true message is hidden
    between what they said.
3. The conversation that reflects the individuals'
    unique traits, such as personality, closeness,
    age difference, and relationship dynamics,
    ensuring it aligns with the provided information
     and avoiding generic dialogue.
4. The conversation should {flow}
5. Conversation does not need to be peaceful, there
    can be arguments, conflict or even curse word.
6. Adjust level of respectfulness with each one's
    emotional state.
7. DO NOT repeat or use similar words that is used to
     describe each character.
8. Avoid repeating what has been said previously in
    the conversation.
9. Use casual words that usually appear in a
    conversation.
10. Format should be "Name (Emotion): Utterance". Do
    not include anything else. Do not use Markdown
    nor double quotes for the utterances.
11. Emotion should fall into the space of Plutchik's
    wheel of emotion.

**Restrictions**:
Avoid using the following or similar terms in the
    conversation
- Trying to help
- I'm doing my best
- Don't come crying to me
- Agree to disagree
- The way I am
```

Template for Summarize the Conversation

```
{conversation}

Summarize the above conversation in 4-5 sentences. It
     should capture the situation and the flow of
    the conversation. It should also indicate the
    outcome of the conversation (i.e., If it ended
    positively or negatively).
```

Template for Self-Verification

```
Use the situation, conversation flow and summary
    given, answer the following tasks. Strictly
    follow the output format.

**Tasks with instruction**:
- Does summary align with both the situation and the
    flow? Respond with only Yes or No.
- On a scale of 1-5, rate the alignment between
    situation and summary. 1 being the summary is
    not describing the situation and 5 being the
    summary is describing the situation. Respond
    with only a number.
- On a scale of 1-5, rate the alignment between
    conversation flow and summary. 1 being the
    summary is not reflecting the flow provided and
    5 being the summary is reflecting the flow.
    Respond with only a number.

Situation: {situation}

Conversation Flow: {flow}

Summary:
{summary}

**Output format**:
Situation: [1-5]
Flow: [1-5]
Overall Alignment: [Yes/No]
```

Template for Norm Violation Discovery and Remediation Suggestion

```
The following is a conversation between {relationship
    }. Explain why this conversation did not go well
     partially or entirely. What are some of the
    norms or rules that are being violated in the
    conversation?

**Information about the participants**:
Participant 1: {person_1_name}, Age {person_1_age}
Participant 2: {person_2_name}, Age {person_2_age}
Relationship: {relationship}
Closeness: {closeness}
How they know each other: {how_they_know}
How long do they know each other: {how_long_they_know
    }

**Requirements**:
1. Norms or rules should be generic meaning it can be
     applied to different scenarios. However, norms
    listed must be applicable to the participants
    given their relationship, closeness, and other
    information provided.
2. Descriptions should describe the norm in a general
     way. Include information like what is the
    expected behavior. Do not mention anything from
    the conversation nor the information provided.
3. Analyze norm violations for both participants.
4. List only one violator for each norm. If both
    participants violated the same norm, list them
    separately.
5. Evidence should be the utterance where the
    violation happened. The utterance must from the
    violator. Only include the utterance without
    name and emotion.
6. Only list the norms that can be observed from the
    text transcript. Do not list norms that require
    further evidence from audio and video. An
    example to avoid is actively listening, as you
    can't observe that from the text.
7. Only list norms that significantly caused this
    conversation to go awry.
8. Suggest a way to make least amount of changes to
    the original utterance that convey the same
    message
 and intention without break the norm violated. The
     suggestion can not be generic. It must align
    with the conversation until that turn and the
    speaker. Do not repect anything said previously.

9. Do not list similar norms. Make sure listed norms
    are common and important.
```

```
10. Do not use Markdown, and follow the output format
     provided. Do not generate anything else.

**Output format**:
Norm:
Description:
Violator:
Evidence:
Suggestion:

**Conversation**:
{conversation}
```

Template for Naturalness

```
Instruction:
Please act as an objective judge and evaluate the
    naturalness of the conversation given. Read the
    conversation between the two participants. Pay
    attention to their dialogue, tone, the flow of
    conversation, as well as whether it resembles a
    typical conversation in the given context.


To evaluate the conversation, first, begin your
    evaluation by providing a short explanation of
    how likely the conversation is going to happen
    as well as an explanation of the naturalness.
    After providing your explanation, you must rate
    the conversation by choosing from the following
    options:
- Rating 5: The conversation sounds entirely natural
    and realistic.
- Rating 4: The conversation sounds reasonably
    natural and flows well.
- Rating 3: The conversation is neither particularly
    natural nor unnatural.
- Rating 2: The conversation lacks some naturalness
    but is not entirely unrealistic.
- Rating 1: The conversation sounds forced, awkward,
    or unrealistic.

Background Information:
Note: If relationship is unknown, try to infer it
    from the conversation given below.
Relationship: {relationship}

Conversation:
{conv}
```

Template for Escalation

```
Objective: Measure the Level of Escalation in a
    Conversation

Instructions:
1. Start by briefly summarizing what happened in the
    conversation. Focus on the key turning points
    and the overall tone.
2. Choose a rating based on how much the conversation
     escalated. Use the following scale:
- Rating 5: The conversation has escalated to a high
    degree of conflict. (Very hostile)
- Rating 4: The conversation has escalated to a
    moderate degree of conflict.
- Rating 3: The conversation has escalated but to a
    low degree of conflict.
- Rating 2: The conversation has not escalated, but
    there is potential for conflict.
- Rating 1: The conversation has not escalated, and
    there is no potential for conflict. (No
    hostility)

Conversation:
{conv}
```

Template for Intervention Conversation Quality

```
Objective: Compare the Overall Conversation Quality
    Between Two Conversations

Instructions:
1. Start by briefly summarizing what happened in each
     conversation. Focus on the key points and the
    overall tone of each conversation.
2. Compare the two conversations based on the
    following criteria:
  - Are the characters in both conversations
      conveying similar messages?
  - Is the core complaint or issue addressed in both
      conversations?
  - Are both conversations about the same situation
      or topic?
3. Choose a rating based on the overall quality of
    the conversations. Use the following scale:
- Rating 5: The conversations are highly similar in
    message, address the same core complaint, and
    are focused on the same situation.
- Rating 4: The conversations are mostly similar in
    message, address mostly the same core complaint,
     and are largely focused on the same situation.
- Rating 3: The conversations share some similarities
     in message, address somewhat the same core
    complaint, and are generally about the same
    situation.
- Rating 2: The conversations have few similarities
    in message, address barely similar core
    complaint, and are loosely related to the same
    situation.
- Rating 1: The conversations are not similar in
    message, address different core complaint, and
    are about different situations.

Conversation 1:
{conv_1}

Conversation 2:
{conv_2}
```

Annotation UI

Correlation between Norm and Relationship

In our study of how relationships influence the nature of norm violations, we systematically categorized conversations based on different relationship types. For each category, we identified tri-grams for norm violations and noted those appearing more than five times.

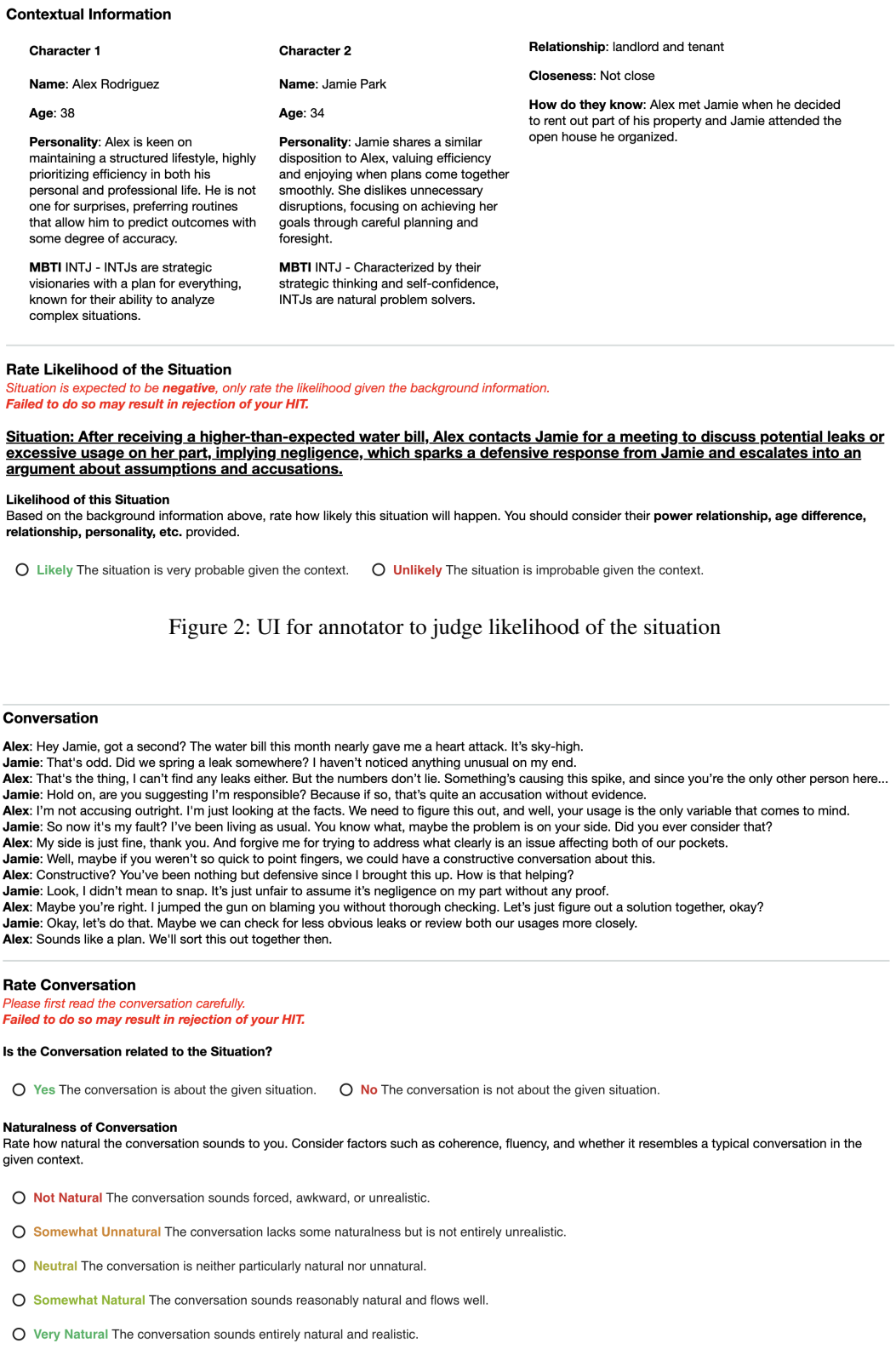

Figure 2: UI for annotator to judge likelihood of the situation

Figure 3: UI for annotator to rate naturalness of the conversation (Below) and whether the conversation aligns with the situation (Above)

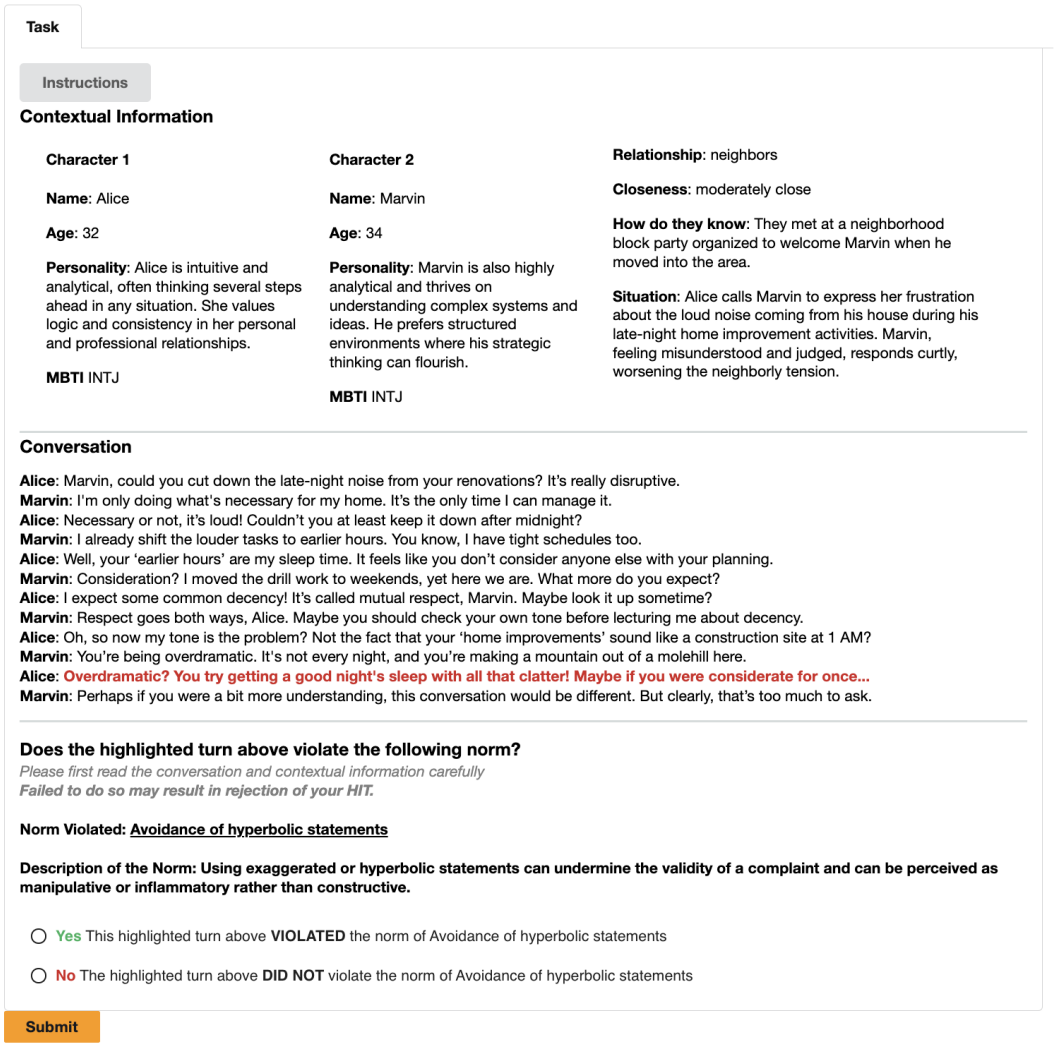

Figure 4: UI for annotator to judge the norm violation

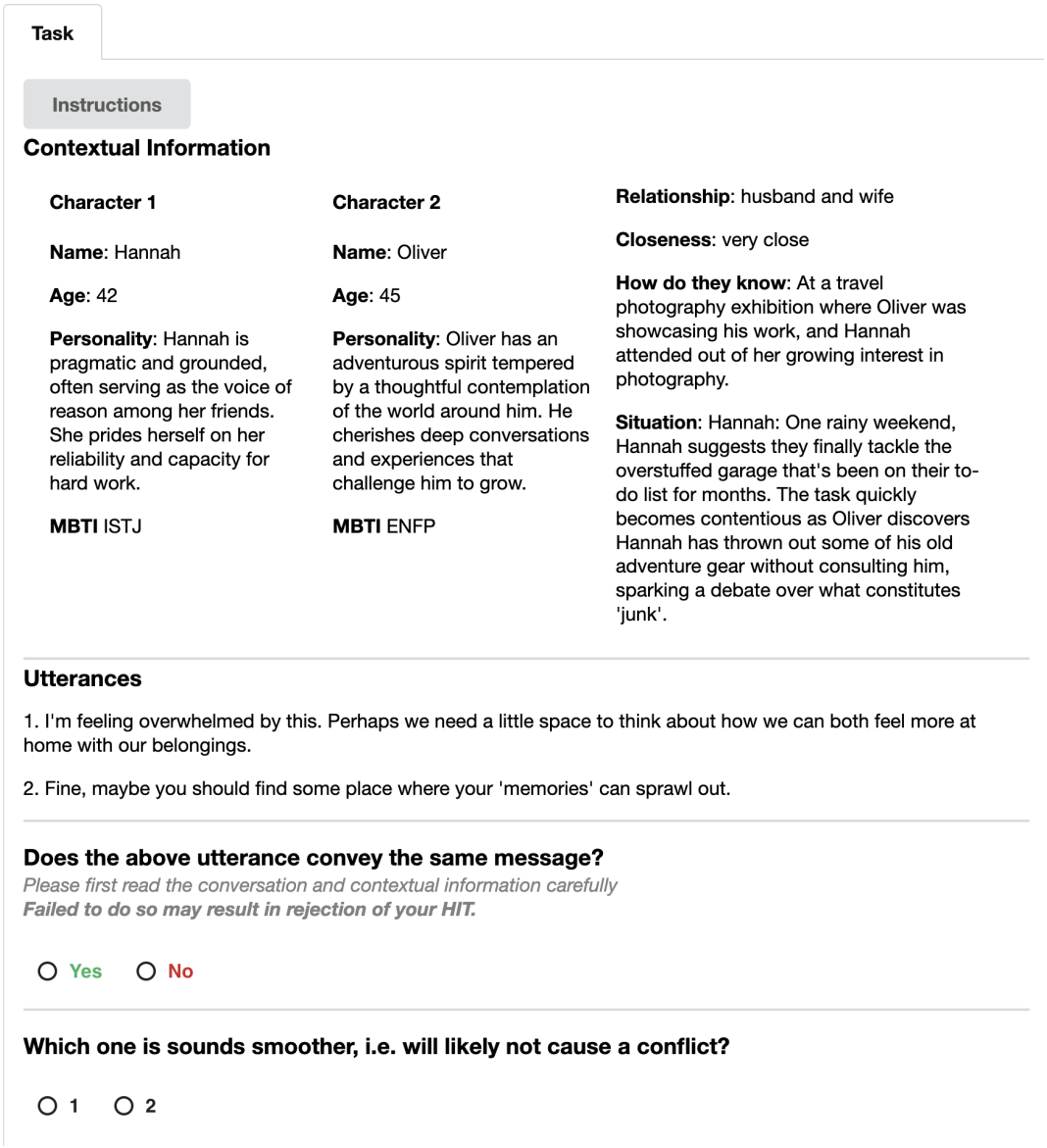

Figure 5: UI for annotator to rate remediation

Upon analyzing each relationship type, we discovered distinct patterns in norm violations that highlight the unique dynamics within each type of relationship.

In `parent-child` relationships, the norms most frequently violated are *maintaining supportive tone* and *respecting personal space*. These findings underline the often hierarchical nature of this relationship, where parents may adopt a more directive tone and impose boundaries, which can lead to emotional and spatial conflicts.

In `sibling` relationships, the most common norm violations include *avoiding accusatory language* and *respecting personal property*. This suggests that siblings, who typically share a more egalitarian and competitive dynamic, are prone to engaging in accusatory exchanges and conflicts over shared or personal items.

When examining romantic relationships, such as `husband-wife` and `boyfriend-girlfriend`, the prevalent norm violations shift to *acknowledging partner's feelings* and *addressing concerns directly*. This finding highlights the expectation for partners to be emotionally supportive and communicative, reflecting the close, intimate nature of such relationships.

In `colleague` relationships, the pattern of norm violations is distinct yet again, with *maintaining a professional tone* and *avoiding personal attacks* being the most common. This is indicative of the professional and often formal environment of the workplace, where maintaining decorum and avoiding personal conflicts is crucial.

These observations reveal that relationship contexts significantly shape the types of norms likely to be violated, capturing the nuanced dynamics specific to each type of interpersonal interaction.

Training Details

Training Arguments

```
==Base Model==
unsloth/Meta-Llama-3.1-8b-Instruct-bnb-4bit

==Lora Args==
r=16
target_modules=[
    "q_proj", "k_proj", "v_proj", "o_proj",
    "gate_proj", "up_proj", "down_proj"
]
lora_alpha=16
lora_dropout=0
bias="none"
use_gradient_checkpointing="unsloth"
random_state=3407
use_rslora=False
loftq_config=None

==Training Args==
per_device_eval_batch_size=32
per_device_train_batch_size=8
gradient_accumulation_steps=2
learning_rate=2e-4
lr_scheduler_type="linear"
warmup_steps=10
num_train_epochs=4
bf16=True
optim="adamw_8bit"
weight_decay=0.01
max_grad_norm=0.3
save_strategy="best"
eval_steps=5,
eval_strategy="steps"
seed=42
greater_is_better=False
metric_for_best_model="eval_loss"
```

Training Prompt Template

```
**Instruction:**
Analyze the given conversation for any violations of
    social norms by either participant. Identify
    each violation, specifying the social norm that
    was violated, the violator, and the turn in
    which the violation occurred. Additionally,
    suggest a more appropriate way for the violator
    to express themselves to avoid escalating the
    situation.

If no social norms were violated, respond with: `No
    clear violation found.` Otherwise, format your
    response as follows:

**Response Format:**
1. Social Norm Violated: [Brief explanation of the
    norm and how it was violated]
Detailed Explanation: [Provide a detailed explanation
     of the violation]
Violator: [Name of the person who violated the norm]
Violated Turn: [Utterance where the violation
    occurred]
Suggestion: [How they could have expressed themselves
     more appropriately]
...

**Conversation:**
{conv}"""
```

Algorithm 1 Conversation Generation Pipeline

Input: Seed relationships R_{pool}
Initialize conversation list: $Conv \leftarrow []$
Initialize participant profiles list: $P \leftarrow []$
Define prompts: $Prompts \leftarrow \{Pt_{pair}, Pt_{sit}, Pt_{conv}, Pt_{qc}, \dots\}$
// Step 1: Generate character profiles for each relationship
for each relationship $r \in R_{pool}$ **do**
    Generate character profile using prompt Pt_{pair}:
        $P \leftarrow P \cup OpenAI(r, Pt_{pair})$
end for
// Step 2: Generate potential situations for each character profile
for each profile $P_i \in P$ **do**
    Generate situations using prompt Pt_{sit}:
        P_i[situations] $\leftarrow OpenAI(P_i, Pt_{sit})$
end for
// Step 3: Filter out similar situations within the same relationship
$P_{filtered} \leftarrow Filter(P, R_{pool})$
// Step 4: Generate conversations and perform quality check
for each profile $P_i \in P_{filtered}$ **do**
    for each situation $s \in P_i$[situations] **do**
        Generate conversation using prompt Pt_{conv}:
            $conv \leftarrow OpenAI(P_i, s, Pt_{conv})$
        Verify conversation quality using prompt Pt_{qc}:
            $isValid \leftarrow OpenAI(conv, P_i, Pt_{qc})$
        if $isValid$ **then**
            Add valid conversation to $Conv$:
                $Conv.add(conv, P_i, s)$
        end if
    end for
end for

Figure 6: Overview of the entire pipeline from a pool of seed relationships to conversation.