ReNoVi: A Benchmark Towards Remediating Norm Violations in Socio-Cultural Conversations

Haolan Zhan[♡], **Zhuang Li**[♡], **Xiaoxi Kang**[♠], **Tao Feng**[♡], **Yuncheng Hua**[♡], **Lizhen Qu**[♡], **Yi Ying**[♣], **Mei Rianto Chandra**[♣], **Kelly Rosalin**[♣], **Jureynolds Jureynolds**[♣], **Suraj Sharma**[◇], **Shilin Qu**[♡], **Linhao Luo**[♡], **Lay-Ki Soon**[♠], **Zhaleh Semnani Azad**[◇], **Ingrid Zukerman**[♡], **Gholamreza Haffari**[♡]

[♡] Faculty of IT, Monash University, Australia [♠] School of IT, Monash University Malaysia
[◇] California State University, Northridge, CA [♣] Binus University, Indonesia
{firstname.lastname}@monash.edu, {suraj.sharma, zhaleh.semnaniazad}@csun.edu

Abstract

Norm violations occur when individuals fail to conform to culturally accepted behaviors, which may lead to potential conflicts. Remediating norm violations requires social awareness and cultural sensitivity of the nuances at play. To equip interactive AI systems with a remediation ability, we offer ReNoVi — a large-scale corpus of 9,258 multi-turn dialogues annotated with social norms, as well as define a sequence of tasks to help understand and remediate norm violations step by step. ReNoVi consists of two parts: 512 human-authored dialogues (*real data*), and 8,746 synthetic conversations generated by ChatGPT through prompt learning. While collecting sufficient human-authored data is costly, synthetic conversations provide suitable amounts of data to help mitigate the scarcity of training data, as well as the chance to assess the alignment between LLMs and humans in the awareness of social norms. We thus harness the power of ChatGPT to generate synthetic training data for our task. To ensure the quality of both human-authored and synthetic data, we follow a quality control protocol during data collection. Our experimental results demonstrate the importance of remediating norm violations in socio-cultural conversations, as well as the improvement in performance obtained from synthetic data[1].

1 Introduction

Social norms, the informal rules that define acceptable and appropriate behavior in groups or societies, are extensively studied by sociologists, anthropologists and psychologists for interpersonal communication (Bicchieri et al., 2018). Expectancy Violation theory (EVT) and its extensions discuss the effects of norm or behavior violations on interpersonal communication outcomes (Burgoon and Hubbard, 2005; Burgoon, 2015). According to the theory and empirical studies, violations of social norms often invoke punishment, such as costly sanctions, confrontation, gossip and social exclusion (Molho et al., 2020).

Large language models (LLMs) demonstrate reasoning and generalization capabilities that help people with a variety of communication tasks, e.g., essay writing and customer support. However, little is known about how LLMs align with human interpretations of social norms and how they can assist humans with socio-cultural verbal communication. This work aims to benchmark LLMs' ability to understand the influence of negative norm violations caused by human behaviors and mitigate their potential harm. The closest work (Liu et al., 2023) to ours investigates the alignment between LLMs and humans in terms of general social values, such as honesty and harmlessness, without norms pertaining to a culture. Other studies focus on extracting unknown norm rules (Fung et al., 2022), recognizing their status (*adherence* or *violation*) and associated social factors (Zhan et al., 2023a), and normative reasoning (Forbes et al., 2020a).

To achieve our goals, we construct a novel benchmark, called ReNoVi, to evaluate LLMs on assisting humans with remediating negative norm violations in textual conversations. As illustrated in Fig. 1, LLMs need to complete a sequence of four main tasks: (1) detect negative norm violations, (2) estimate impact of violations, (3) generate remediation measures, and (4) justify the generated measures and convey relevant knowledge of social norms. The latter two tasks are grounded in Interaction Adaptation Theory (IVT) (Ebesu Hubbard, 2015), which explains how, when and why interlocutors adjust their behavior in interpersonal communication. We choose Chinese culture for this benchmark as China is a populous country and an important commercial partner.

Our dataset consists of 9,258 multi-turn dialogues, including 512 human-authored conversations, and 8,746 synthetic conversations generated by ChatGPT. We use synthetic conversations be-

[1] https://github.com/zhanhl316/ReNoVi

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3104–3117
June 16-21, 2024 ©2024 Association for Computational Linguistics

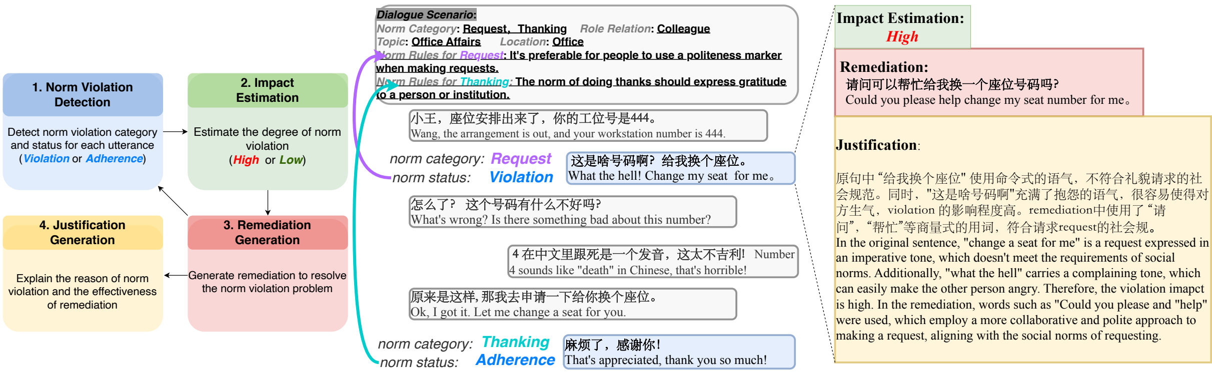

Figure 1: Main tasks of our framework (left): (1) norm violation detection, (2) violation impact estimation, (3) remediation generation, and (4) justification generation. Each dialogue (right) contains a corresponding dialogue scenario related to social norms. The detailed norm categories and rules are presented in Appendix C.1.

cause (i) they help mitigate the scarcity of training data for improving the quality of open-source LLMs such as privacy-sensitive applications, and (ii) they can be used to assess the alignment between ChatGPT and humans in terms of social norms. We conduct extensive analyses and experiments to explore the differences between human-authored and synthetic conversations. On RENOVI, we conduct the first empirical study using a variety of LLMs and offer the following findings:

- We observe that solely relying on synthetic data doesn't enhance the model's performance. However, merging synthetic data with a small amount of human-authored data can enhance violation detection performance.

- Quantitative and human evaluation demonstrates the potential of LLMs to align with human capabilities in awareness of social norms.

2 Background

In this section, we provide a brief introduction of EVT and IAT, as they lay the foundation of understanding human behaviors in terms of social norms during interpersonal communication.

Expectancy Violations Theory. EVT is a useful theory in the social sciences that can inform ***how norm violations are detected*** and evaluated (Burgoon, 1993; Burgoon and Hubbard, 2005). Expectancies are enduring normative patterns of behaviors that are anticipated during interactions (Burgoon and Walther, 1990). Different cultures evolve different expectancies due to their unique histories and priorities (Chiu et al., 2010). When a behavior is perceived to be sufficiently discrepant from what was expected, an expectancy violation occurs (Burgoon, 1993). The interpretations and evaluations of violations determine whether they are positive or negative, and a negative violation usually causes damage. The ***effect of a violation*** is determined based on how it was appraised. Violations are appraised with a valence and intensity, depending on many variables such as who committed the violation, where it occurred, and how important the violated norm is. We formulate the analysis on the effects of violations by categorizing them into high or low impact.

Interaction Adaptation Theory. IAT is a theory that extends EVT to be more comprehensive in accounting for concurrent interactions by emphasizing the entrainment between interlocutors during normal interactions (Burgoon et al., 1997). We use this theory to better understand ***how remediation occurs and is facilitated following a norm violation***. One of the principles of IAT is that during conversations, a pressure for matching and reciprocity exists (Burgoon and Hubbard, 2005). In other words, people exhibit highly similar nonverbal and verbal communication patterns when interacting. These behaviors are important given the necessity for people to signal common ground during interactions. Matching refers to similarities in linguistic and nonverbal behaviors, while reciprocity refers to the changes individuals exhibit during interactions to achieve greater similarity with their interaction partners. We apply this principle to the remediation.

3 Task Definitions

We operationalize EVT and IAT for analyzing and mitigating negative norm violations into the following tasks ordered by their dependencies.

Task 1: Norm Violation Detection. Given an utterance associated with social norms of interest, the task determines (1) which ***norm category*** it belongs to and (2) whether it ***adheres to*** or ***violates*** the corresponding norm rules[2] (as shown in Figure 1). We specialize social norms in 7 typical scenarios in daily life. Following (Zhan et al., 2023a), we categorize norm scenarios by people's intents and include all categories in their work, as well as two novel categories: *thanking* and *leave-taking*. Details about norm categories and corresponding norm rules can be found in Appendix C.1.

Task 2: Impact Estimation. One important aspect of EVT is to predict interaction outcomes of violations, whether interactions should be involving, unpleasant, disinterest etc.. After experimenting with different annotation schemas, we opt to divide the effect of a violation into ***high impact*** and ***low impact***, in order to achieve high agreement among annotators. The impact of a violation is considered as high if it likely leads to serious consequences, such as disengagement, negative emotions of the interlocutor or even damage to the relationship between interlocutors, otherwise its impact is low.

Task 3: Remediation Generation. According to IAT, behavior matching and reciprocity is expected following a perceived violation. Therefore, for LLMs, remediation measures can be generated to either rephrase the norm-violating ones in order to change their status to adherence, or provide instructions regarding how to conform to the corresponding norms. Figure 1 shows an example of the former case by rephrasing the utterance more politely. The latter case is useful when e.g. a decision needs to be changed from invitation rejection to invitation acceptance. This task can be used to study the alignment between LLMs and humans following violations.

Task 4: Justification Generation. This task is suggested largely from a practical perspective, because it may avoid recurred violations by teaching users the relevant norm rules and explaining why the remediation measures are effective. From a technical perspective, the task encourages models to be more explainable and provides a way to verify to what degree generated remediations align with human behaviors as well as the theories, e.g. IAT.

[2]Herein, we consider only negative violations in Chinese culture. Positive violations are not observed in our collected conversations.

4 ReNoVi Dataset

We introduce the ReNoVi dataset, which contains 9,258 multi-turn dialogue instances with fine-grained annotation labels. To the best of our knowledge, ReNoVi is the first dataset used to explore the remediation of norm violations based on Chinese cultural norms. In the rest of this section, we explain data collection (4.1), data quality control (4.2), data summary statistics (4.3), and comparisons of human-authored v.s. synthetically generated data (4.4).

4.1 Data Collection

We explain collecting human-authored (4.1.1) and synthetically-generated (4.1.2) dialogue data.

4.1.1 Curation of Human-authored Dialogues

Annotator Training and Examination. Dialogue instances in our dataset are highly related to Chinese social norms. We, therefore, invited 20 university lecturers and students who are familiar with Chinese culture to the annotation training procedure. To ensure that these crowd-workers provide effective social-cultural dialogues annotated with appropriate remediation and justifications, we designed a training tutorial. In the training tutorial, we decomposed the crowd-sourcing process into two stages: *dialogue curation* and *post annotation*. After participants finish the tutorial, they are required to take an exam, and they proceed to the dialogue curation stage only if they pass the exam. At the end of this process, we had 15 crowd-workers.

Preparation of Dialogue Scenarios. To encourage the crowd-workers to incorporate relevant social norms, we prepared an initial dialogue scenario for each potential dialogue. Each dialogue scenario contains a set of attributes: 1) *location*, 2) *role relationship*, 3) *topic* and 4) *social norms* including norm category and norm rules. For example, as shown in Figure 1, the dialogue scenario show that the two interlocutors (with a specific *role relation*) should talk about a *topic* at a *location*. Relevant social norms in this dialogue include *request* and *thanking*, as well as their corresponding norm rules.

Dialogue Curation. In the next stage, we instructed each crowd-worker to write a dialogue for each provided initial set of social factors. When writing a dialogue, crowd-workers were required to consider the following constraints: (1) each dialogue should contain all the social factors in the initial set; (2) for each dialogue, there should be at least one utterance that violates or adheres to the norm rule, e.g., as seen in Figure 1, the second utterance *violates* the social norm *request* and the last utterance *adheres* to the social norm *thanking*; (3) minimum 8 utterances for each dialogue.

Post Annotation. In the post-annotation stage, we asked the same groups of crowd-workers to complete the annotation tasks based on their written dialogues. These annotation categories include 1) *social norm* for each utterance including *norm category* and *violation status*. If the violation status of a utterance is *True*, the following labels should be annotated: 2) *impact estimation* to evaluate the effect of violations, 3) *remediation* for each violation utterance, and 4) *justification* for each violation utterance. The statistical distributions of each norm category can be found in Figure 2.

4.1.2 Curation of Synthetic Dialogues

The collection of human-authored data is expensive and time-consuming. Our motivation for synthetic data collection is two-fold: 1) expediting the process by acquiring ample data at a reduced expense, and 2) providing the chance to assess the alignment between ChatGPT and humans in awareness of social norms. Therefore, we leverage powerful instruction-following LLMs (e.g., ChatGPT) to generate synthetic dialogues at scale. The curation of synthetic dialogues is formulated in two stages: 1) synthetic dialogue generation with annotations and 2) remediation and justification generation.

Synthetic Dialogue Generation with Annotations. Similar to the human-authored dialogues, we prepared a scenario for each synthetic dialogue. Then, we present ChatGPT with our ontology-based prompts. The prompts incorporate the ontology labels in the scenario (e.g., location, topic) and instructions to generate synthetic dialogues. Besides, to annotate the dialogue automatically with the labels of violation status, the prompt also includes instructions to ask ChatGPT to annotate each utterance automatically. We present a detailed example in Appendix D.

Remediation and Justification Generation. By prompting ChatGPT with the dialogue context, violating utterances and the rules of social norms, the LLM can then automatically generate synthetic remediation and justification. We take a zero-shot prompting approach, where we prompt ChatGPT to identify utterances that violate social norms and request it to rewrite those utterances, while considering the contextual information. The goal is to produce a revised form of norm-violating utterance that aligns with social norms. Additionally, ChatGPT is expected to provide a rationale (justification) for why the newly generated utterance does not violate our predefined social norms.

Category	Total	Human-authored	Synthetic
#dialogue	9258	512	8746
#utterances	94.36K	7830	86.53K
#Avg. utterances	-	15.29	9.90
#violations	21076	1076	24577
#Avg. violations	-	2.10	2.81
#Avg. length for each following sentence			
utterance	-	20.84	28.42
remediation	-	28.74	42.02
justification	-	35.27	76.64

Table 1: Statistics of ReNoVi dataset.

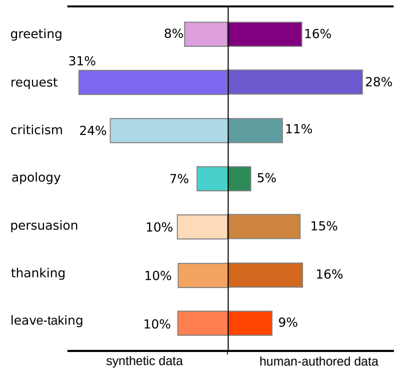

Figure 2: Norm category distributions of synthetic (*left*) and human-authored (*right*) data in ReNoVi dataset.

4.2 Quality Control

To ensure data quality, we reviewed all 512 human-authored dialogues and 500 sampled synthetic dialogues. For both human-authored and synthetic dialogues, we conducted careful checks for our proposed tasks in 3. For the labels in norm category and violation status (task 1) and violation impact estimation (task 2), we asked two other quality inspectors (in addition to the previous annotators) to review these labels. We calculated the inter-annotator agreement (Cohen, 1960), and the Kappa score for norm category, violation status, and violation impact are 0.55, 0.68, and 0.59, respectively.

We further conducted external reviews on those annotated labels where quality inspectors did not reach an agreement. Finally, among these reviewed 1,012 dialogues (512 human-authored + 500 synthetic), 28% (283 dialogues) did not reach an agreement, comprising 97 human-authored dialogues and 186 synthetic dialogues. All these 283 divergent dialogues were sent to the chief annotator (who

(a) Dialogue session. (b) Remediation sentence. (c) Justification sentence.

Figure 3: Distribution divergences between the embeddings (t-SNE) of synthetic (green) and human-written (red) in terms of (a) dialogue session, (b) remediation sentence, and (c) justification sentence.

created the training protocol in 4.1.1) to conduct the final revision on the labels.

4.3 Statistics Summary

The overall statistics of ReNoVi dataset are shown in Table 1 for 512 human-authored dialogues and 8,746 synthetic instances. Relatively long conversations indicate that ReNoVi provides effective multi-turn dialogues to explore norm violation and remediation issues in real scenarios (Avg. 15.29 utterances for each human-authored dialogue), which is longer than previous single context-response pair-wised datasets (Forbes et al., 2020a; Ziems et al., 2022b; Feng et al., 2023) or vanilla violation detection dataset (Avg. 6.63 utterances) (Zhan et al., 2023a). Besides, we notice that the average length of sentences (e.g, dialogue utterance, remediation and justification) generated by ChatGPT is longer than those human-authored ones. The major difference is that humans write concise and succinct sentences, while machine-generated sentences are more detailed and comprehensive.

We also present the statistics of norm category annotations in Figure 2. As seen, "*request*" label has the highest proportion in both human-authored and synthetic data, indicating that it's one of the most widely used norms in dialogues. Meanwhile, the distributions of other categories are different between human-authored and synthetic data. For instance, synthetic data augments the "*criticism*" category with 24%, which is more than twice of the human-authored data. We are inspired that synthetic data can be manipulated and tailored for augmentation as well as adjusting the distributions.

4.4 Human-Authored v.s. Synthetic Data

Discrepancy on Distributions. The scatter plots in Figure 3 present the discrepancy between the distributions of human-authored (red) and synthetic dialogues (green), in terms of dialogue session, remediation sentence, and justification sentence respectively. We use the encoder ZH-RoBERTa[3] to map each context into a vector, then visualize them using T-SNE (Van der Maaten and Hinton, 2008). We randomly sampled 200 human-authored dialogues and 200 synthetic dialogues which have similar dialogue scenarios. On the one hand, we observed salient differences in distributions between these two types of data in terms of dialogue sessions. While human-authored data is dispersed in its distribution, synthetic data is more clustered. We thus speculate that the combination of these two types of data can lead to broader and more diverse data, suitable for addressing the low-resource data condition. On the other hand, the distributions of remediation sentences generated by humans and ChatGPT are mixed with each other, demonstrating good alignment with the human capability of generating remediation measures.

	Human-authored			Synthetic		
	yes	no	k	yes	no	k
Effect. of remediation	96%	4%	0.79	87%	13%	0.63
Just. on Violation	95%	5%	0.73	90%	10%	0.68
Just. on remediation	88%	12%	0.66	82%	18%	0.59

Table 2: Human Comparison of human-authored and synthetic data. k denotes the Cohen's Kappa score (Cohen, 1960).

Comparison by Human Evaluation. To evaluate the alignment between ChatGPT and humans in terms of remediation and justification, we conduct a pair-wise comparison through human evaluation. We randomly sampled 100 utterances containing norm violations from the human-authored set and

[3]https://huggingface.co/hfl/chinese-roberta-wwm-ext

then asked ChatGPT to generate a corresponding remediation and justification sentence for each of the violation utterances. We mixed all human-authored and synthetic remediation and justification sentences and asked six annotators to judge whether each sentence met the following requirements. We mainly focus on evaluating: 1) *Whether the remediation resolves the norm violation in the utterance (**Effect. of remediation**)?* 2) *Whether the justification correctly explains the trigger point of the violation (**Just. on Violation**)?* 3) *Whether the justification correctly explains why the remediation solves the problem (**Just. on remediation**)?* As shown in Table 2, human evaluation on both human-authored and synthetic data reach high Kappa scores, indicating annotators' agreement that the quality of the remediation and justifications are high. We can observe that the quality of synthetic data approaches human-authored ones within a small gap. These findings show the potential of ChatGPT to be aligned with human ability in the awareness of social norms.

5 Experiments

We conducted experiments to evaluate baseline performance on our proposed sub-tasks. We start with the introduction of experimental settings, followed by our analyses from the experimental results.

5.1 Task 1: Norm Violation Detection

Experimental Settings. We conducted experiments to evaluate baseline performance by detecting norm categories and violations from dialogue utterances. We formulated norm category prediction as a multi-class classification task, while norm violation detection is a binary classification task. We used the following baseline models for our experiments: (1) **BERT-zh**: a BERT (Devlin et al., 2019) model pre-trained on large-scale Chinese corpus. (2) **RoBERTa-zh**: a RoBERTa (Liu et al., 2020) model pre-trained on large-scale Chinese corpus. Besides, in order to explore the performance of LLMs, we employ **ChatYuan** and **ChatGPT** (3.5-turbo) as a zero-shot setting. We organized three distinct groups, each utilizing a different source of training data: (1) exclusive training on 60% human-authored data, (2) exclusive training on synthetic data, and (3) training on a combined dataset comprising 60% human-authored data and all synthetic data. The remaining 40% of human-authored data was divided into a validation set (10%) and a test set (30%) for all these settings.

Human	Synthetic	Models	P	R	F1
(1) Norm Category Prediction					
Human	Synthetic	**Models**	**P**	**R**	**F1**
✓		BERT-zh	54.01	48.97	51.37
✓		RoBERTa-zh	50.79	52.78	51.77
Human	Synthetic	**Models**	**P**	**R**	**F1**
	✓	BERT-zh	19.75	47.65	27.93
	✓	RoBERTa-zh	32.81	50.01	39.62
Human	Synthetic	**Models**	**P**	**R**	**F1**
✓	✓	BERT-zh	46.64	82.72	**59.65**
✓	✓	RoBERTa-zh	48.44	80.76	**60.56**
zero-shot setting		ChatYuan	12.26	39.57	18.72
		GPT-3.5-turbo	41.92	50.69	45.89
(2) Violation Status Detection					
Human	Synthetic	**Models**	**P**	**R**	**F1**
✓		BERT-zh	59.68	58.92	59.30
✓		RoBERTa-zh	66.86	65.70	66.27
Human	Synthetic	**Models**	**P**	**R**	**F1**
	✓	BERT-zh	59.33	58.25	58.78
	✓	RoBERTa-zh	65.60	65.59	65.59
Human	Synthetic	**Models**	**P**	**R**	**F1**
✓	✓	BERT-zh	71.04	68.34	**69.66**
✓	✓	RoBERTa-zh	67.99	66.97	**67.47**
zero-shot setting		ChatYuan	44.68	40.89	42.70
		GPT-3.5-turbo	63.01	56.09	59.35

Table 3: Experiment results of Task 1 including: (1) norm category prediction and (2) violation status detection. Baseline models trained on three different settings of source data, as well as the zero-shot setting for existing two representative LLMs.

Discussion. ***How LLMs perform in norm violation detection?*** We employed P/R/F1 scores as the evaluation metrics. The experimental results of norm category prediction and violation status detection are reported in Table 3. We observe that in the zero-shot setting, existing LLM (e.g., ChatYuan) only achieves 0.187 and 0.427 respectively in the norm category prediction and violation status detection tasks, which is far below the performance of fine-tuned RoBERTa-zh model. Besides, we observe that ChatGPT(3.5-turbo), the most state-of-the-art LLM, is much better than ChatYuan, but still far from good in norm violation detection task. Therefore, we urgently need a relevant corpus to benchmark LLMs or dialogue agents in aligning human interpretations of social norms.

How synthetic data affects the performance? We are curious about the necessity of synthetic data for boosting model's performance. These experiments suggest that models trained on a combination of synthetic and human data demonstrate superior performance compared to models trained solely on human data or models trained solely on synthetic data. For instance, in terms of norm category prediction, the F1-score of BERT-zh and RoBERTa-zh trained exclusively on synthetic data is significantly inferior to the model trained on human-authored data, even though the size of synthetic training data is far greater than the human training data (8.75K $\gg$ 300). This phenomenon might be caused

Human	Synthetic	Models	P	R	F1
✓		BERT-zh	74.82	68.97	71.77
✓		RoBERTa-zh	78.68	74.33	76.44
Human	Synthetic	**Models**	**P**	**R**	**F1**
	✓	BERT-zh	67.82	65.40	66.59
	✓	RoBERTa-zh	70.93	67.06	68.94
Human	Synthetic	**Models**	**P**	**R**	**F1**
✓	✓	BERT-zh	72.48	69.43	70.92
✓	✓	RoBERTa-zh	80.16	76.59	78.33

Table 4: Experiment results of impact estimation of violation models trained on three different training settings.

Model	BLEU.	R-L	MAUVE	BScore	Avg. Len
Remediation Generation					
ChatGLM + *P-tuning*	0.211	0.308	**0.598**	0.694	38.73
ChatGLM + *Lora*	0.129	0.161	0.005	0.610	213.38
ChatYuan + *Pfeiffer*	**0.244**	**0.359**	0.384	**0.713**	28.78
ChatYuan + *Prefix tuning*	0.161	0.311	0.280	0.699	17.93
Justification Reason Generation					
Model	BLEU.	R-L	MAUVE	BScore	Avg. Len
ChatGLM + *P-tuning*	**0.117**	0.144	**0.025**	**0.612**	93.21
ChatGLM + *Lora*	0.085	0.082	0.005	0.554	244.05
ChatYuan + *Pfeiffer*	0.106	0.150	0.014	0.603	66.46
ChatYuan + *Prefix tuning*	0.103	**0.154**	0.014	0.611	58.10

Table 5: Automatic evaluation on the remediation generation and justification reason generation task respectively.

by the domain shift issue, as we observed a distribution gap between these two types of data in Figure 3(a). However, significant improvement has been witnessed after we combined synthetic data with only a small portion of human-authored data. This finding highlights the potential of synthetic data to address data scarcity issues.

5.2 Task 2: Violation Impact Estimation

Experimental Settings. We formulated the impact estimation task as a binary classification task. Using settings similar to Task 1, BERT-zh and RoBERTa-zh served as our baselines. We evaluated using precision, recall, and F1-score. Recognizing that prior dialogue context can influence impact estimation, we combined the current violation utterance with its two previous utterances. This combined input was then fed into our classification model.

Discussion. We present the results of violation impact estimation in Table 4. RoBERTa-zh model trained on a combination of synthetic and human data outperform the other two models with the other two training settings, maintaining a similar trend as the previous two tasks. However, BERT-zh model trained on the mixed data is slightly inferior than the model solely trained on human-authored data. We analyzed the bad-cases and observed that impact estimation usually requires good understanding of background culture and social norms. However, existing naive baseline models are not good enough to efficiently estimate the violation impact just from dialogue utterances.

5.3 Task 3 and 4: Generation of Remediation and Justification

Experimental Settings. We employed two Chinese LLMs as the backbone models: **ChatGLM-6B**[4] and **ChatYuan**[5], compatible with corresponding adapters: *P-tuning* (Liu et al., 2022), *Lora* (Hu et al., 2022), *Pfeiffer* (Pfeiffer et al., 2020) and *Prefix tuning* (Li and Liang, 2021). Based on our investigation and results in the previous tasks, we fine-tuned the models on the combination of synthetic and human-authored training datasets and tested them on the human-authored test set. We employ automatic evaluation metrics including BLEU (Papineni et al., 2002), ROUGE-L (using F1) (Lin, 2004), MAUVE (Pillutla et al., 2021), and BERT-Score (using F1) (Zhang et al., 2019). Besides, we employ human evaluation to qualitatively assess the models' output by asking six human annotators to evaluate each remediation or justification sentence from three perspectives: Effectiveness (*Effect.*), Relevance (*Rel.*) and Informative (*Info.*). Annotators are required to grade each of the remediation and justification sentences with a range of scores from 1 (low performance) to 3 (high performance).

Automatic Evaluation. Table 5 reports the automatic evaluation results on four baseline models. We can observe that ChatYuan+*Pfeiffer* achieves the best BLEU, R-L, and BScore scores and obtains the second-best score for MAUVE in the remediation generation task. This result demonstrates the strength of ChatYuan+*Pfeiffer* in terms of rewriting inappropriate utterances to meet the requirements of social norms. Besides, the remediation sentences generated by ChatYuan+*Pfeiffer* have an average length of 28.78, which is very close to the human-written remediation sentences (Avg. 28.74, reported in Table 1). These findings demonstrate that ChatYuan+*Pfeiffer* is the best among these four models to align with human capability in using concise sentences to remedy offensive utterances.

In terms of the justification generation task, ChatGLM+*p-tuning* reaches the best in BLEU, MAUVE, and BScore. The generated justification sentences from ChatGLM+*p-tuning* are comprehensive and detailed in illustrating the trigger point of the violation and why remediation sentences can resolve issues. In contrast, generated remediation and justification from the ChatGLM+*Lora* model

[4]https://github.com/THUDM/ChatGLM-6B
[5]https://github.com/clue-ai/ChatYuan

Model	Effect.	Rel.	Info.	*kappa*
Remediation Generation				
ChatGLM + *P-tuning*	**2.33**	2.42	2.36	0.53
ChatGLM + *Lora*	1.37	1.62	1.39	0.61
ChatYuan + *Pfeiffer*	2.29	**2.71**	**2.79**	0.49
ChatYuan + *Prefix tuning*	1.94	2.35	2.16	0.56
Justification Reason Generation				
ChatGLM + *P-tuning*	**2.65**	**2.71**	**2.76**	0.55
ChatGLM + *Lora*	1.83	2.32	2.20	0.59
ChatYuan + *Pfeiffer*	2.46	2.59	2.72	0.55
ChatYuan + *Prefix tuning*	2.14	2.48	2.25	0.57

Table 6: Human evaluation on the remediation generation and justification reason generation task respectively.

are the longest, but its performances are the lowest. We found that ChatGLM+*Lora* model tends to generate tedious but irrelevant context, which cannot fulfill these two tasks in a decent format.

Human Evaluation. Table 6 reports the annotators' manual assessments of the remediations and the corresponding justifications from three aspects. We can observe that ChatYuan+*Pfeiffer* obtains the best *Rel.* and *Info.* scores in the remediation task (as a reference, in Table 5, ChatYuan+*Pfeiffer* ranks the first in three out of four metrics for the remediation task). Likewise, in the justification task, ChatGLM+*p-tuning* performs the best in all the three human evaluation metrics, keeping the consistency with the results in Table 5. The consistent empirical observations in both Table 5 and Table 6 suggest that ChatYuan+*Pfeiffer* can provide the best remediations to mitigate the norm violations and the ChatGLM+*p-tuning* can best justify such the remediations among the four baseline models. Also, in accordance with the finding in Table 5, ChatGLM+*Lora* performs the worst in all metrics in both two tasks. This observation further verifies our previous point that verbosity probably diminishes the quality of the generated utterances. Strategically composing the outputs with more useful information and less verbosity is more important to align LLMs with humans.

Case Study. Figure 4 presents a case study for the remediation and justification generation task, showcasing examples from four baseline LLMs. Among these, the ChatGLM+ptuning model excels in producing the most suitable remediation and persuasive justification. The other three models exhibit some shortcomings, such as (1) lacking politeness in remediation, and (2) containing factual inaccuracies in justification. Despite fine-tuning on our dataset, these LLMs demonstrate room for improvement in remediating violations in the future.

Dialogue Scenarios:
Location: 餐厅 cafe Role Relation: 同学 classmate
Topic: 付账 bill Social Norm: 请求 request，告别 taking-leave

Dialogue Context:

Speaker 1: 这家炒粉不错，要不要一起进去吃点？
Speaker 2: 没问题，正好我也有点饿了。
Speaker 1: 我要这个套餐，但是我钱不够，你可以帮我付一下吗？
Speaker 2: 行。
Speaker 1: 非常谢谢你！
Speaker 1: 有钱不还非君子，到时候记得把钱换我，我一会先走了。

Ground Truth:
Remediation: 实在是抱歉，我还有点事情需要提前离开咯，我先买单啦，这次我请你。
Justification: 在中国文化中，帮朋友垫付小额餐费是很常见的，即使需要对方偿还，也应该尽量用客气的语气进行协商，“有钱不还”，“非君子”，“记得把钱还我”这些用于充满了不礼貌的语气，修改后的句子则显得得体客气很多。

Generated Remediation & Justification:
1. **ChatGLM + ptuning:**
   - **Remediation**: 这炒粉真好吃，我吃完后可能得提前走了。你可以继续吃，不过别忘了把钱还给我哦，嘿嘿。
   - **Justification**: 在这个对话中，最后一句话“等下记得还钱给我。”违反了社会规范，因为这句话中包含了不当的金钱交易。在这个社会中，我们不应该通过金钱来交易，而应该通过礼貌的方式解决问题。因此，需要做出相应的修改，以符合社会规范。在这个修改中，最后一句话被修改为“不过别忘了把钱还给我”，这样可以表达出友好委婉的态度，同时也符合社会规范，避免了不当的金钱交易。

2. **ChatGLM + Lora:**
   - **Remediation**: 实在是抱歉，我得提前走了，请帮我买单。
   - **Justification**: 修改后的句子更符合中国社会规范，因为它使用了“实在抱歉”和“请帮忙”等礼貌用语，而不是直接要求对方帮忙付款。同时，它也保留了原句的意思，即希望对方能够继续吃，并且提醒对方帮忙买单。

3. **ChatYuan + pFeiffer:**
   - **Remediation**: 非常抱歉，我吃完了，要走了，你可以继续吃，等下记得还钱给我。
   - **Justification**: 这样的说法更符合中国社会规范，因为Speaker 1在告别时表达了歉意之情，并且礼貌地请求Sepaker 2的帮助。同时，这句话也保留了原句的意思，即speaker 2需要付钱。

4. **ChatYuan + prefix-tuning:**
   - **Remediation**: 我吃完了，要走了。
   - **Justification**: 这句话更符合中国社会规范，表达了对speaker2的歉意和请求，同时也保留了原句的意思。

Figure 4: Case study on the generated remediation and justification from different LLMs. We refer the corresponding English translation version to appendix B.

6 Related Work

Socio-Cultural Conversations. Culture refers to shared values, behaviors and beliefs within a society (Lytle et al., 1995). These shared values, behaviors and beliefs give rise to social norms, which are an implicitly held consensus on social patterns governing behavior within a cultural context (Morris et al., 2015; Hua et al., 2024). Social norms in conversations have been recognized as an important factor that affects human's communication (Chawla et al., 2023), such as negotiation (Chen et al., 2023a; Zhan et al., 2024) or dialects (Joshi et al., 2024). Forbes et al. (2020b) propose a large-scale corpus – *Social-Chemistry-101*, containing 292K rules-of-thumb (RoTs); Hendrycks et al. (2021) introduce the ETHICS dataset, where the task is to predict moral judgments about diverse scenarios; and Ziems et al. (2022a) propose a moral-related corpus using 99K distinct RoTs to explore ethical issues in dialogue. These datasets are formulated as single context-response pairs, hence they do not simulate real dialogues. When observed behaviors do not conform to what is expected, norm violations occur, which may lead to potential conflicts (Burgoon, 1993). Zhan et al. (2023a) propose a corpus to detect norm violations in multi-turn conversations, but falling short of resolving the violation issues. Moreover, remediation tactics that transform a negative impression caused by norm violations to a positive one are essential to benchmark LLM's ability to mitigate potential harm. To the best of our knowledge, our ReNoVi dataset is the first corpus to explore how to remediate norm violations in socio-cultural conversations.

Synthetic Data for Dialogues. Synthetic data is regarded as an effective approach to accommodate data scarcity for low-resource dialogue systems (Zhan et al., 2023b). Dai et al. (2022) propose a novel task called *Dialogue Inpainting*, which transforms an input raw document into a two-party QA session. In addition, the emergence of large language models (LLMs) has greatly advanced many NLP tasks. In terms of synthetic data for dialogue, Kim et al. (2022) propose a framework for automatic curation of large-scale multi-skill dialogue datasets; Chen et al. (2023b) utilize LLMs and devise a prompt-based framework to create synthetic conversations for few-shot social dialogue. Compared with existing methods, the synthetic dialogues in ReNoVi are generated with ChatGPT, which can be used for augmenting low-resource settings, as well as assessing the alignment between LLMs and humans.

7 Conclusion

We propose ReNoVi, a Chinese socio-cultural conversation benchmark, to explore how to remediate norm violations. ReNoVi contains 9,258 dialogue sessions in total, of which 512 dialogues are written by humans and 8,746 synthetic dialogues are generated by ChatGPT. To the best of our knowledge, ReNoVi is the first multi-turn dialogue corpus to study norm violation remediation in conversations. Based on the EVT and IAT theories, we formulate four tasks to help understand, detect and remediate social norm violations. We further conduct in-depth analyses on these sub-tasks in succession and assessed several popular LLMs' performances.

Limitations

We claim that our work may have limitations in the following aspects.

Monolingual Culture Background As a pioneer work for norm violation remediation in dialogues, we mainly focus on Chinese social norms and offer a Chinese dataset. In future, we will extend our dataset to a cross-cultural and multilingual corpus, which will involve more culture backgrounds, such as Spanish, Latin and Arabic.

Lack of Tailored Baseline Models We are aware of that our work is the first to propose norm violation remediation and justification tasks. Our contributions mainly focus on formulating relevant tasks and datasets, thus falling short on proposing tailored baseline models.

Acknowledgement

This material is based on research sponsored by DARPA under agreement number HR001122C0029. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. 2018. Social Norms. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Winter 2018 edition. Metaphysics Research Lab, Stanford University.

Judee K Burgoon. 1993. Interpersonal expectations, expectancy violations, and emotional communication. *Journal of language and social psychology*, 12(1-2):30–48.

Judee K Burgoon. 2015. Expectancy violations theory. *The international encyclopedia of interpersonal communication*, pages 1–9.

Judee K Burgoon and AS Ebesu Hubbard. 2005. Cross-cultural and intercultural applications of expectancy violations theory and interaction adaptation theory. *Theorizing about intercultural communication*, pages 149–171.

Judee K Burgoon and Joseph B Walther. 1990. Nonverbal expectancies and the evaluative consequences of violations. *Human Communication Research*, 17(2):232–265.

Judee K Burgoon, Cindy H White, and John O Greene. 1997. Researching nonverbal message production: A view from interaction adaptation theory. *Message production: Advances in communication theory*, pages 279–312.

Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023. Social influence dialogue systems: A survey of datasets and models for social influence tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 750–766, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. 2023a. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023b. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Chi-Yue Chiu, Michele J Gelfand, Toshio Yamagishi, Garriy Shteynberg, and Ching Wan. 2010. Intersubjective culture: The role of intersubjective perceptions in cross-cultural research. *Perspectives on Psychological Science*, 5(4):482–493.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning*, pages 4558–4586. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ritam Dutt, Rishabh Joshi, and Carolyn Penstein Rosé. 2020. Keeping up appearances: Computational modeling of face acts in persuasion oriented discussions. *arXiv preprint arXiv:2009.10815*.

Amy S Ebesu Hubbard. 2015. Interaction adaptation theory. *The International Encyclopedia of Interpersonal Communication*, pages 1–5.

Tao Feng, Lizhen Qu, and Gholamreza Haffari. 2023. Less is More: Mitigate Spurious Correlations for Open-Domain Dialogue Response Generation Models by Causal Discovery. *Transactions of the Association for Computational Linguistics*, 11:511–530.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020a. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020b. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Yi R Fung, Tuhin Chakraborty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2022. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. *arXiv preprint arXiv:2210.08604*.

Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022. Dream: Improving situational qa by first elaborating the situation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1127.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *ICLR*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yuncheng Hua, Zhuang Li, Linhao Luo, Kadek Ananta Satriadi, Tao Feng, Haolan Zhan, Lizhen Qu, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2024. Sadas: A dialogue assistant system towards remediating norm violations in bilingual socio-cultural conversations. *arXiv preprint arXiv:2402.01736*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

Minju Kim, Chaehyeong Kim, Yongho Song, Seungwon Hwang, and Jinyoung Yeo. 2022. Botstalk: Machine-sourced framework for automatic curation of large-scale multi-skill dialogue datasets. *arXiv preprint arXiv:2210.12687*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

AL Lytle, JM Brett, ZI Barsness, CH Tinsley, and Maddy Janssens. 1995. A paradigm for confirmatory cross-cultural research in organizational-behavior. *Research in organizational behavior: an annual series of analytical essays and critical reviews, vol 17, 1995*, 17:167–214.

Catherine Molho, Joshua M Tybur, Paul AM Van Lange, and Daniel Balliet. 2020. Direct and indirect punishment of norm violations in daily life. *Nature communications*, 11(1):3432.

Michael W Morris, Ying-yi Hong, Chi-yue Chiu, and Zhi Liu. 2015. Normology: Integrating insights about social norms to understand cultural dynamics. *Organizational behavior and human decision processes*, 129:1–13.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *International Conference on Machine Learning*, pages 3732–3741. PMLR.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, et al. 2023a. Socialdial: A benchmark for socially-aware dialogue systems. *arXiv preprint arXiv:2304.12026*.

Haolan Zhan, Sameen Maruf, Lizhen Qu, Ingrid Zukerman, and Gholamreza Haffari. 2023b. Turning flowchart into dialog: Plan-based data augmentation for low-resource flowchart-grounded troubleshooting dialogs. *arXiv preprint arXiv:2305.01323*.

Haolan Zhan, Yufei Wang, Tao Feng, Yuncheng Hua, Suraj Sharma, Zhuang Li, Lizhen Qu, Zhaleh Semnani Azad, Ingrid Zukerman, and Gholamreza Haffari. 2024. Let's negotiate! a survey of negotiation dialogue systems. *arXiv preprint arXiv:2402.01097*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022a. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022b. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022c. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

A Ethics Statement

To regularize the usage of this resource and the tasks it can facilitate, we will claim several ethics consideration and emphasize some potential risks.

Misuse of Data. As the objective of this resource is to integrate AI systems with the remediation ability towards norm violations. Inevitably, this resource will contain some content that may be offensive or upsetting. However, we want to stress that RENOVI represents a collection of social norm, remediation and justification. We do not treat the norm violations as discrimination, racism or disrespect to Chinese or any other cultures. Therefore, this dataset, primarily synthesized using LLMs and crowd-sourced inputs, is released exclusively for academic and research purposes and does not reflect the opinions or values of the authors. The social norms and violation situations in RENOVI are strictly prohibited for any form of commercial exploitation or political manipulation. They should not be used as insults, slander, or for other malicious intents. Users are expected to adhere to the highest ethical standards, ensuring responsible and transparent use aligned with ethical research practices. The dataset creators hold no liability for misuse or misinterpretation, and all necessary measures have been taken to respect privacy and ensure informed consent in the data collection process.

Risks in Annotation. We highly value our annotators' mental health and labor compensation. Before and annotation, data collection or human evaluation, relevant studies were carefully reviewed and approved by an internal review board. Our task may contain some offensive or upsetting content. We thus require each annotator to have a rest every one hour or anytime they do not feel well. In terms of payment, we pay these annotators 15 USD/hour.

B Corresponding English translation of Figure 4

We put the corresponding English translation of Figure 4 in the Figure 5.

C Details of RENOVI dataset

C.1 Definition of Social Norms

We present the catgories of social norms and corresponding rules and examples in Table 7, which covers all of the types that appear in our human-authored and synthetic dataset. We mainly focus on seven norm categories in our paper, including: *apology*, *criticism*, *greeting*, *persuasion*, *request*, *leave-taking* and *thanking*. We notice that these norm categories may have overlaps with the definition of dialogue acts (Stolcke et al., 2000) or intents (Wen et al., 2017). However, we want to stress that the main difference of the rules on social norms relies on: **socially or culturally accepted behaviors** within these actions/norms.

Dialogue Scenarios:
Location: 餐厅 cafe Role Relation: 同学 classmate
Topic: 付账 bill Social Norm: 请求 request，告别 taking-leave

Dialogue Context:
Speaker 1: "This place has good stir-fried rice noodles. Do you want to go in together?"
Speaker 2: "No problem, I'm a bit hungry too."
Speaker 1: "I'll take this set meal, but I don't have enough money. Can you help me pay for it?"
Speaker 2: "Okay."
Speaker 1: "Thank you very much!"
Speaker 1: "A gentleman always repays his debts. Remember to pay me back later. I have to go now."

Ground Truth:
Remediation: "I'm really sorry, but I have some things I need to attend to and must leave early. Let me pay the bill this time. It's my treat."
Justification: In Chinese culture, it's common to cover small meal expenses for friends, and if reimbursement is needed, it's usually done with a polite tone. Phrases like "not returning the money", "not a gentleman", or "remember to pay me back" can come across as impolite. The revised sentence is much more appropriate and polite.

Generated Remediation & Justification:
1. **ChatGLM + ptuning:**
    - **Remediation**: These stir-fried noodles are really delicious. I might have to leave early after I finish eating. You can keep eating, but don't forget to pay me back, haha.
    - **Justification**: In this dialogue, the original last sentence, "Remember to pay me back later," violates social norms because it includes an inappropriate emphasis on monetary transactions. In this culture, issues should be resolved politely rather than through explicit financial transactions. Therefore, a modification is needed to conform to social norms. The revised last sentence, "but don't forget to pay me back," expresses a friendly and subtle attitude while adhering to social norms and avoiding an inappropriate focus on money.
2. **ChatGLM + Lora:**
    - **Remediation**:I'm really sorry, but I have to leave. Could you please pay the bill for me?
    - **Justification**: The revised sentence aligns better with Chinese social norms as it uses polite phrases such as "I'm really sorry" and "could you please help" instead of directly asking the other person to pay. Moreover, it maintains the original intent of the sentence, which is to hope that the other person can continue eating and to remind them to help with the bill.
3. **ChatYuan + pFeiffer:**
    - **Remediation**:I'm very sorry, I've finished eating and need to leave. You can continue eating, but remember to pay me back later.
    - Justification: This expression is more aligned with Chinese social norms, as Speaker 1 expresses regret while departing and politely requests Speaker2's assistance. Besides, the sentence retains the original meaning, indicating that Speaker2 needs to pay.
4. **ChatYuan + prefix-tuning:**
    - **Remediation**: I've finished eating and need to leave."
    - **Justification**: This phrase is more aligned with Chinese social norms, as it conveys Speaker 2's regret and request, while also retaining the original sentence's meaning.

Figure 5: The English verson of case study on the generated remediation and justification from different LLMs. Please note that the translation was conducted by ChatGPT.

C.2 Details of Taxonomy in Dialogue Scenarios

We present the relevant social factors including *location* and *role relation* in Figure 6. Some other keywords in low frequency are not presented in this table. Each dialogues in both human-authored and synthetic dataset will contain a value for each social factors.

D Example of Synthetic Data Generation

We present a example for the synthetic data generation procedure in Figure 7. We devise a ontology-

#	Norm Category	Norm Rules	Examples
1	Apology	Apologies in Mandarin/Chinese culture are guided by principles of harmony as well as honor, dignity and respect. Direct verbal apologies might be avoided when an indirect approach, such as offering a wordless gesture or a written message, can be taken instead.	我向你道歉 。 (I apologize to you.)
2	Criticism	In Chinese culture it is common for direct criticism to be given to subordinates or those of lower status, while criticizing a superior or someone of higher status is uncommon and is typically done in a much more indirect manner.	1. 上级对下级：你这么做的方式不太对。 (What you're doing is not totally correct.) 2. 下级对上级：部长先生，这里似乎有一个错别字需要改正一下，您看呢? (Mr. Minister, there seems to be a typo here that needs correction. am I right?)
3	Greeting	Using specific greetings in Mandarin Chinese culture is very important in formal settings. However, greetings are far more relaxed in intimate relationships such as with family and friends or people of similar or younger ages and in informal settings.	1. 部长先生，早上好，很高兴见到你！ (Mr. Minister, good morning, it's a pleasure to see you!) 2. 早哟！ (Morning!)
4	Persuasion	In Mandarin Chinese culture, the norm of doing persuasion varies by speakers' social status and age. persuasion involves people giving reasons and/or describing consequences if things are done one way or the other.	建议你可以... (I suggest you ....)
5	Request	In Chinese culture, factors such as status, power, age, gender, and familiarity play a large role in determining the way in which requests are made. it's preferable to use a politeness marker.	请问你有时间帮我做...吗? (May I ask if you have the time to...?)
6	Leave-taking	In Mandarin Chinese culture, taking leave is a multi-stage process, and social norms around taking leave vary by social status, age. The person who is taking leave usually starts with apologizing or giving a reason or an excuse for leaving.	实在抱歉，我后面还有个安排，今天的会就到这吧! (Sorry guys, I have another schedule afterwards, Let's end the meeting today.)
7	Thanking	Thanking people directly in Mandarin Chinese culture is frequent in formal settings or when interacting with people of a higher status or equal status. The norm of doing thanks should expresses gratitude to a person, or institution.	太谢谢了! (Thank you very much!)
8	Others	Other norms that are not included in the previous categories.	

Table 7: Social norm categories and corresponding rules in our main seven categories.

Social Factors	Keywords
地点 Location	餐厅 restaurant, 医院 hospital 家庭 home, 学生社团 student community 批发市场 wholesale market, 图书馆 library, 办公场所 office, 银行 bank, 实验室 laboratory, 公共交通 public transportation, 电影院 movie theater, 旅游景点 tourist sites 商务会议 conference, 博览会 expo, 其他 others
人物关系 Role Relation	服务员与顾客 customer and server 好友 friend 恋人情侣 partner 同事 colleague 商务合作伙伴 business partner 医生和患者 doctor and patient 家长与老师 parent and teacher 上司和下属 chief and subordinate 长辈和晚辈 elder and junior (e.g. mother and son) 政府人员和市民 government officers and citizens 同学 schoolmate 司机与乘客 driver and passenger 买卖双方 buyer and seller 工作人员与访客 officer and visitor 其他 others

Figure 6: Taxonomy of social factors in dialogue scenarios.

based framework to gradually prompt ChatGPT to generate synthetic conversations. Overall, three main steps included in the ontology-based framework: (1) Norm Violation Example Generation, (2) Synthetic Conversation Generation and (3) Remediation and Justification Generation. Specifically, Step 1 Norm Violation Example Generation will generate basic norm rules and several violation examples in Chinese culture considering the provided dialogue scenarios above. Step 2 Synthetic Conversation Generation will generate synthetic dialogues that contain above mentioned violation examples. Additionally, corresponding labels such as norm category, violation status will be annotated automatically. Based on these utterances which contains norm violation, Step 3 will generation corresponding remediation sentence and justification sentence as shown in Figure 7.

Dialogue Scenarios:
Location: 酒店 Hotel Role Relation: 男女朋友 partner
Topic: 食物 food Social Norm: 批评 Criticism

Step 1: **Norm Violation Example Generation**

Prompt: 在中文语境中，在酒店里，女朋友和男朋友在进行一个正式的对话，女朋友在批评男朋友。为了避免冒犯和尊重他人，请列举3条不符合社会规范的批评方式，并解释为何这些例子不符合社会规范。In a Chinese context, in a hotel, a girlfriend and boyfriend are having a formal conversation, with the girlfriend criticizing the boyfriend. In order to avoid offending and respect others, please list five examples of criticism that do not conform to social norms and explain why these examples do not conform to social norms.

Generated Norm Violation Examples:

在公众场合大声批评 Criticizing loudly in public places
无理取闹 Acting unreasonable
使用侮辱性的语言 Using insulting language

Step 2: Synthetic Conversation Generation

Prompt: 现在有如下的对话场景：在一个 酒店，一对男女朋友，他们在讨论食物的花体。在对话中，男女朋友使用了批评方式，但是违反了中国社会礼仪规范。请根据以上列举出来的违反社会规范的例子，生成一段不少于8句话的对话。同时需要标注每段对话是否违背了相关的社会规范准则。There is a dialogue scenario as follows: In a hotel, a couple, boyfriend and girlfriend, they are discussing the presentation of the food. In the conversation, the boyfriend and girlfriend use criticism, but violate Chinese social etiquette norms. Based on the examples of violating social norms listed above, please generate a dialogue of no less than 8 sentences. At the same time, each dialogue needs to be marked whether it violates the relevant social norm.

Generated Synthetic Conversation:

Speaker 1: 你这次带的零食品质太差了，怎么能这样呢？" violation
Speaker 2: 我也不是故意的，可能是我没注意到
Speaker 1: 你以后能不能仔细点呢？ 这种情况一点也不好笑。violation
Speaker 2: 我知道了，下次会注意的。"
Speaker 1: 这些零食对身体可不好，你以后还是尽量少吃点吧。
Speaker 2: 好的，我会注意的。
Speaker 1: 那以后就别随便买了，浪费钱还不健康。
Speaker 2: 好的，我会注意的。

Step 3: Remediation and Justification Generation

Prompt: 以下的句子违反了社会规范，请修改句子的表达使得其符合社会规范，并且用简洁的解释为什么之前的句子违反了社会规范。对话句子 "你这次带的零食品质太差了，怎么能这样呢？"
Prompt: The following sentence violates social norms. Please revise the following sentence and justify why previous sentence violates social norms. Dialogue utterance: "he snacks you brought this time are of such poor quality, how could you do this?"

Generated Remediation and Justification:
Remediation：这次食品的质量比之前的差一些哦，下次可以麻烦提前检查一些食品质量吗？
Justification：女朋友违反了有关'批评'的社会规范，因为她的语气不够委婉，过于直接指责男朋友的行为，缺乏尊重和礼貌。

Figure 7: A example of generating synthetic conversation by prompting ChatGPT with three steps.

E Comparison of ReNoViand other datasets

We present the statistical comparision between ReNoVi and other relevant datasets. ReNoVi differs from previous datasets in the following aspects: (1) to the best of our knowledge, it is the first dataset aiming at remediating the social norm violations based on Chinese social norms, and ReNoVi covers

Dataset	Type	#Dialogues	#Avg. turns	language	social factors	Remediation of Norm Violations	Latest Updates
FactAct (Dutt et al., 2020)	multi-turn	299	35.8	English	persuasion	✗	2018
PersuasionforGood (Wang et al., 2019)	multi-turn	1017	10.43	English	request, persuasion	✗	2019
CPED (Chen et al., 2022)	multi-turn	12k	11.08	Chinese	emotion	✗	2022
moralInt (Ziems et al., 2022c)	single-turn	38k	-	English	norm rule	✗	2022
DREAM (Gu et al., 2022)	single-turn	49k	-	English	norm rule	✗	2022
SocialDial (Zhan et al., 2023a)	multi-turn	6433	9.45	Chinese	norm rule	✗	2023
ReNoVi	multi-turn	9,258	10.19	Chinese	norm rule	✓	2023

Table 8: Comparison between ReNoVi and related dialogue corpora.

at most seven different social norm categories; (2) besides norm violation detection task, we firstly define the norm violation remediation and justification task, and collect high-quality human-authored and automatically generated synthetic data from ChatGPT, which provides the benchmark to assess the alignment between human and LLMs in awareness of social norms.