LLM-Human Pipeline for Cultural Grounding of Conversations

Rajkumar Pujari and Dan Goldwasser
Purdue University, West Lafayette, USA
{rpujari,dgoldwas}@purdue.edu

Abstract

Conversations often adhere to well-understood social norms that vary across cultures. For example, while *addressing work superiors by their first name* is commonplace in the Western culture, it is rare in Asian cultures. Adherence or violation of such norms often dictates the tenor of conversations. Humans are able to navigate social situations requiring cultural awareness quite adeptly. However, it is a hard task for NLP models.

In this paper, we tackle this problem by introducing a *Cultural Context Schema* for conversations. It comprises (1) conversational information such as emotions, dialogue acts, etc., and (2) cultural information such as social norms, violations, etc. We generate ∼110k social norm and violation descriptions for ∼23k conversations from Chinese culture using LLMs. We refine them using automated verification strategies which are evaluated against culturally aware human judgements. We organize these descriptions into meaningful structures we call *Norm Concepts*, using an interactive human-in-the-loop framework. We ground the norm concepts and the descriptions in conversations using symbolic annotation. Finally, we use the obtained dataset for downstream tasks such as emotion, sentiment, and dialogue act detection. We show that it significantly improves the empirical performance.

1 Introduction

Social norms define behavioral expectations shared across groups and societies (Sherif, 1936). They can help explain the differences in the way people from different cultural backgrounds react to the same situation (Triandis et al., 1994; Finnemore, 1996). Hymes (1972) describe *norms of interaction* as 'shared rules that implicate the belief system of a community', capturing the importance of representing norms when, either human or AI-systems, attempt to make sense of social interactions from different cultures.

Motivated by Large Language Models emergent reasoning abilities (Wei et al., 2022a; Srivastava et al., 2023; Camburu et al., 2018; Suzgun et al., 2022) several recent works attempted to create repositories of cultural norms using novel prompting approaches followed by automated verification of the generated descriptions (Fung et al., 2023; Li et al., 2023; Ziems et al., 2023). However, these efforts had limitations such as using synthetic conversations or focusing on a handful of situations. LLMs also tend to suffer from several reliability issues such as hallucinations (Rawte et al., 2023) or high sensitivity to prompt structure (Kojima et al., 2023; Prystawski et al., 2023; Wei et al., 2022b).

These limitations motivate a more general and principled approach for *Cultural Context Understanding*, which, we argue, should be viewed as a pragmatic reasoning task. Norms are situated in specific settings and are associated with the social roles participants play (Hare, 2003). Different expectations can be associated with individuals based on attributes such as their status, profession, or gender. Furthermore, these expectations are situation-dependent, for example, changing when engaging in professional or leisure activities.

To that end, in this paper, we propose a novel Cultural Context Grounding framework for conversations. We tackle three key problems. First, *norm representation* capturing the multi-party situated social expectations. Second, norm induction, i.e., populating the suggested representation with norm concepts, emerging from conversational data, and associating them with relevant contextual information. Finally, grounding the norm concepts in conversational data at scale and creating a dataset of norm-schema aligned conversations.

Our norm representation solution is inspired by the notion of *scripts* (Schank and Abelson, 1977), i.e., structured representations of expected activities for different roles relevant to a specific scenario, in our case mapping to social scenarios. Unlike past

work, Ziems et al. (2023), that developed a schema representing expectations over situated *actions*, our goal is to capture how social norms manifest in conversational behavior. To allow for pragmatic inferences mapping the norm definition to conversations, we intentionally separate between *factual components*, capturing observed information about the settings and content of the conversation, and *cultural norm components*, capturing the expected behaviors and impact of violating that expectation. Fig. 1(a) depicts this separation.

Our framework follows a multi-stage approach to obtaining cultural information for existing conversations and grounding this information in conversation-specific details. We focus mainly on relevant social norms and their violations within these conversations. We then evaluate the usefulness of the created cultural context dataset both qualitatively and quantitatively. Fig. 2 presents an overview of our framework. We describe it in §4.

We leverage LLMs such as GPT-3.5 to generate a large but potentially noisy corpus of conversation-specific cultural information. Then, we leverage an *interactive human-in-the-loop* process (Alg. 1) that efficiently utilizes culturally proficient human annotation to organize this information into meaningful concepts. We further ground the generated descriptions in the conversation details, such as participants, their relationships, and their behavior using symbolic annotation. Following this, we experiment with automated verification strategies to filter the obtained cultural information at scale. Then, we organize the conversations and the obtained cultural information into a meaningful schema structure. Finally, we propose a neural graph schema model that leverages obtained data to improve the empirical performance on conversation understanding tasks such as emotion detection, sentiment detection, and dialogue act detection across multiple datasets. Overall, our contributions can be summarized as follows:

1. We propose a novel Cultural Grounding pipeline for conversation understanding using LLM & culturally-aware human annotation.
2. We introduce *Norm Concepts* and employ a human-in-the-loop (HiL) framework to create human-validated concepts supported by data.
3. We leverage automated verification strategies to clean the LLM-generated data. We further symbolically ground the conversations in norm concepts. We create a high-quality, large-scale dataset for cultural understanding.
4. We leverage the cultural schema information to improve downstream conversational task performance. We present meaningful visualizations and human evaluation experiments to showcase the quality.[1]

2 Related Work

Social Norm Datasets: A few recent works have leveraged LLMs such as GPT-3 to create useful cultural norm datasets using structure prompting approaches (Fung et al., 2023; Li et al., 2023; Ziems et al., 2023). While these works either focus on a small set or synthetically generated conversations, we generate $63k$ norm descriptions for $23.5k$ real conversations and ground them using a principled cultural grounding pipeline.

Social Grounding: Pacheco et al. (2023) propose an interactive concept discovery for COVID-19 tweets. Several works address the tasks of principled grounding in social domain(Smith et al., 2018; Roberts et al., 2019; Roy and Goldwasser, 2021; Demszky et al., 2019; Pujari et al., 2024). However, they mainly focus on social grounding in political settings. We focus on the cultural aspect of social context which is challenging to obtain explicit data.

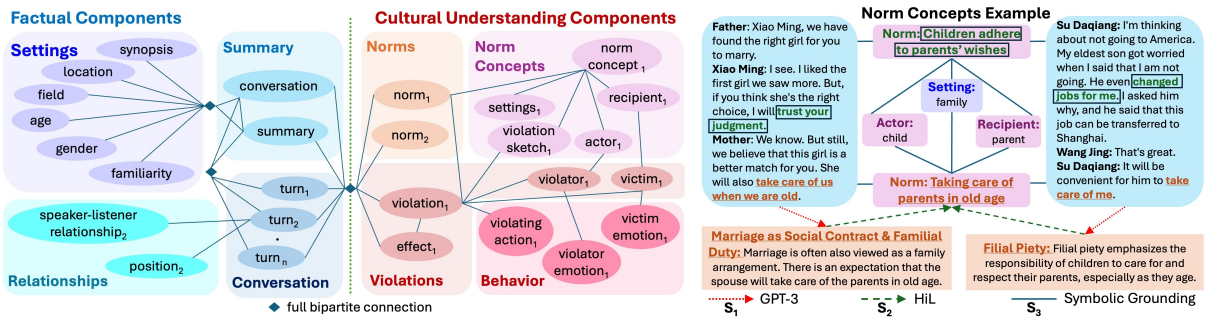
Automated Verification using LLMs: Several previous works address automated annotation using LLMs and refining their generations.(Chiang and Lee, 2023; Li et al., 2021; Bano et al., 2023; Bang et al., 2023; Yao et al., 2022; Hendrycks et al., 2020). We build upon the frameworks proposed by Wu et al. (2023) and Arabzadeh et al. (2024) to operationalize our annotation framework.

3 Data

Dataset: MPDD	
Language: Chinese **Description:** Conversations from Chinese TV series scripts	**# Convs:** 4,141 **# Turns:** 25,546 **Tasks:** Emotion
Dataset: CPED	
Language: Chinese **Description:** Textual dataset with multi-modal features from 40 Chinese TV shows	**# Convs:** 11,832 **# Turns:** 132,723 **Tasks:** Dialogue Act, Emotion, & Sentiment
Dataset: LDC CCU ZH Text	
Language: Chinese **Description:** Chinese textual dataset containing online forum and chat conversations	**# Convs:** 6,763 **# Turns:** 98,821 **Tasks:** Norm Violation, Emotion, & Dialogue Act

Table 1: List of our raw data sources. ZH: Chinese; #Convs: Conversations; #Turns: Conversation Turns

[1]Code and Data: https://github.com/pujari-rajkumar/cultural-schema-naacl2025



(a) Proposed Schema Structure

(b) *Norm Concepts & Symbolic Explanations* example. S_* denotes relevant pipeline stages from Fig. 2.

Figure 1: Proposed *Cultural Context Schema Structure for Conversations* with an example instance

We build upon three existing dataset: MPDD (Chen et al., 2020), CPED (Chen et al., 2022), and LDC CCU Chinese Text[2] datasets. These datasets consist of Chinese conversations annotated with emotions, sentiments, dialogue acts, and social norm violations. Each dataset is annotated for some of these tasks, but not all. Detailed statistics of the dataset are presented in Tab.1. We describe each dataset briefly below.

MPDD Dataset: (Chen et al., 2020) This dataset contains 4,141 dialogues from Chinese TV series scripts. It is annotated with emotions, listeners, and speaker-listener relationships for each turn. The dataset contains 25,546 conversation turns.

CPED Dataset: (Chen et al., 2022) CPED contains transcripts from 40 Chinese TV shows. It contains 11,832 conversations. It is annotated for emotions, sentiments, and dialogue acts. This dataset also provides multi-modal features such as face position, etc. We don't use them in this work.

LDC CCU ZH Text Dataset: This dataset is a conversational understanding dataset from the DARPA CCU program. This dataset consists of text, audio, and video data in several languages. We use the text portion of the Chinese dataset for our experiments. It contains annotations for emotion, dialogue act, social norm violation status, and conversation change points. We focus on emotion and dialogue act identification tasks. This dataset also has metadata about settings, age of participants, familiarity, etc, annotated. The dialogue act annotation for this dataset is not complete. The annotation is focused only on a pre-determined set of dialogue acts. We deal with this by using *other* class for unlabeled data.

[2] https://www.darpa.mil/program/computational-cultural-understanding

As not all datasets have all components annotated, we train the Llama-3.1-70b-Instruct (Dubey et al., 2024) model to predict missing fields for each dataset using QLoRA (Dettmers et al., 2024) fine-tuning. We predict relationships for CPED and LDC datasets. We predict LDC-style metadata for CPED and MPDD datasets.

4 Cultural Context Grounding

Our high-level objective is to '*conceptualize & operationalize cultural understanding in conversational behavior*.' We identify three key steps to make progress towards this objective:

1. Formalize relevant cultural context for better conversational understanding.
2. Obtain high-quality cultural context data at scale, leveraging native-culture expertise and ground the conversations in this context.
3. Evaluate the obtained cultural context dataset for *correctness* and *usefulness*.

First, we propose a graph-based schema structure for culturally enriched conversational representation. Our schema consists of two complementing segments: *factual* components and *cultural understanding* components. We present an overview of the schema structure in Fig. 1 and a detailed description of the schema structure in §4.1.

Then, we introduce a robust pipeline for obtaining cultural information for real conversations and grounding the conversations in it. We efficiently leverage (1) native-culture human expertise, (2) LLMs as knowledge elicitors, (3) LLMs as symbolic annotators, and (4) LLMs as multi-agent verifiers, in this pipeline. We obtain a large-scale corpus for ∼23k Chinese conversations from three existing datasets (Tab. 1). We present an overview of the proposed pipeline in Fig. 2 and a detailed

pipeline description in §4.2.

Finally, we evaluate (1) *correctness* of the cultural information obtained using elicitor & symbolic annotator LLMs, and (2) *usefulness* of the schema for conversational understanding. We measure *correctness* against human annotations (§5). We evaluate the *usefulness* of cultural schema data via conversational tasks such as emotion, sentiment, and dialogue act detection (§6). We present full statistics of our pipeline data collection process in Tab. 7 in appendix H.

4.1 Schema Structure

Context often dictates whether or not a specific behavior is considered normative. For example, while patronizing younger people might be commonplace in families in some cultures, it could be considered offensive in a professional setting. Especially if the younger person is a work superior. On the other hand, even within the same social setting, norms might vary circumstantially. While joking about an elder's forgetful nature might be okay in some situations, it could be considered insensitive if they are suffering from dementia.

This guides us toward two distinct genres of information that could influence conversational behavior: factual and cultural information. To account for this, we propose a schema structure with two distinct segments (Fig. 1(a)). The *factual* segment of our schema consists of settings, summary, conversation, & relationship components. *Cultural Understanding* segment consists of social norms, violations, norm concepts, & behavior components.

While factual information such as age group, location, and relationships is available for many datasets, descriptions of relevant social norms, their violations, and how they affect the conversation are harder to procure. We address this challenge by efficiently using LLMs and human annotation.

4.2 Grounding Pipeline

Grounding in NLP usually refers to linking text or speech to real-world concepts such as entities, attitudes, etc (Chandu et al., 2021). Conversational grounding can encompass various aspects, such as mutual beliefs, shared knowledge, assumptions, and so on (Clark, 1996). In this work, we focus specifically on the *cultural knowledge* aspect of conversational understanding.

Obtaining exhaustive knowledge of a culture's social norms at scale is impractical. Therefore, we propose a bottom-up approach. We use real-world conversations to mine situation-dependent social norms, leveraging LLMs as *knowledge elicitors*. Native-culture human annotators then create structured abstractions over these descriptions, which we refer to as *norm concepts*. Building upon Pacheco et al. (2023), we devise an interactive framework that amplifies human judgments and scales them to the entire dataset. Then, we use LLMs as *symbolic annotators* to ground conversations in human-generated norm concept structures.

As LLMs are susceptible to hallucinations and bias, we evaluate the obtained dataset and symbolic annotations against human judgments. Then, we further employ LLMs as multi-agent verifiers to significantly improve the quality of the dataset. We present an overview of the proposed pipeline approach in Fig. 2. We further discuss details of each step of the pipeline in the following subsections.

4.2.1 LLM Description Generation

Two major challenges we face in building our cultural context dataset are (1) collecting structured knowledge about social norms as they manifest in real-world conversations and (2) scaling the descriptions of culturally nuanced behaviors observed in conversations. We address these challenges by leveraging existing conversational datasets and LLMs as *knowledge elicitors*, respectively. We provide raw conversations to the LLM and instruct it to explain observed behavior from a cultural perspective. For each conversation, we obtain descriptions of relevant social norms, potential violations, and their effects on the conversation trajectory and the participants. We present an example of norm descriptions in Fig. 1(b). We further provide the exact prompts and a full example of LLM outputs in the appendix D. We expect this step to be noisy as LLMs are susceptible to hallucinations. However, our primary focus is to obtain broad coverage of diverse cultural nuances that influence behavior in conversations. We evaluate the quality of LLM generations in §5.

4.2.2 HiL Norm Concept Discovery

As noted in Clark (1996), humans actively draw from a *'common ground'* when engaging in conversations. From a cultural standpoint, this common ground includes shared cultural beliefs. Humans often encounter social interactions where cultural awareness is practiced, which makes them adept at *situating* conversations in *cultural common ground*. Many aspects of this cultural common ground are

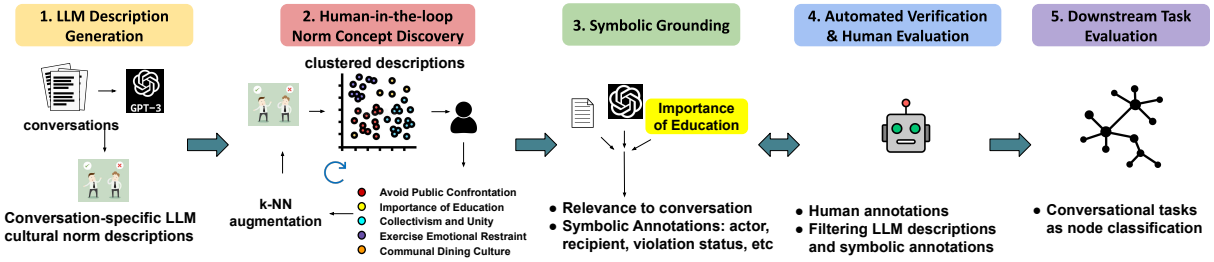


Figure 2: Proposed Cultural Context Grounding Pipeline for Conversations

highly general and serve us in a variety of situations. However, structured datasets that serve as a common ground for NLP models are hard to create solely using either human annotation or automated methods. Hence, in an attempt to create such a resource at scale, we employ culturally aware humans in an interactive human-in-the-loop (HiL) framework, which amplifies their judgments in a semi-automatic fashion.

Algorithm 1 Interactive Norm Concept Discovery

Input: Conversations and their norm descriptions
Outputs: Discovered *norm concepts* with *symbols*, many-to-one mapping: $\langle$desc, conv$\rangle \rightarrow$ concept

1: Cluster norm descriptions using k-means.
2: Cultural experts create *norm concepts* by selecting 5-10 samples of closely related descriptions and providing symbolic structure.
3: Perform k-NN augmentation of unmapped norm descriptions to each norm concept.
4: Experts inspect augmented clusters and mark 5-10 good & bad examples for each concept.
5: Re-assign norm descriptions to concepts using good & bad cluster centers.
6: Go back to step 1 with the remaining unmapped norm descriptions.

Among LLM-generated cultural data, we observe overlapping norm descriptions across conversations, with minor variations. More interestingly, we also find closely related descriptions that can be grouped under the same theme. Consider the example in Fig. 1(b). Norm descriptions '*marriage as a familial duty: spouse is expected to take care of parents in old age*' and '*filial piety towards parents in old age*' are both defined by the common theme '*care of parents in old age as child's responsibility*.'

To capture such themes in a principled manner, we introduce ***norm concepts***. They are abstractions over cultural beliefs which influence several related behaviors. We associate them with symbolic explanations, thus creating structured representations.

A norm concept is characterized by an activation setting, a violation sketch, and actor and recipient roles. *Actor* role describes people expected to adhere to the norm concept. *Recipients* perceive/experience the consequences of adherence or violation. An example concept, *'Respect for Authority'*, is presented in Tab. 2. The symbolic structure of norm concepts is shown in Fig.1(b). The goal of HiL methodology is to support the discovery process via interaction with data.

Concept Name: Respect For Authority
Description: Respecting hierarchies in family, professional, & organizational settings. It involves individuals respecting the decisions, suggestions, orders, and advice from those in higher positions
Settings: workplace, family, organizations
Violation Sketch: Behavior that intentionally contradicts the expectations and decisions of the people in charge
Actors: sub-ordinates or people in an inferior social position such as students, children, etc
Recipients: people in a position of power or authority over other people

Table 2: Cultural expert annotation of symbolic structure for discovered norm concept *Respect for Authority*

Pacheco et al. (2023) propose an interactive concept learning framework for tweets. We extend their framework for norm concept discovery. We outline our process in Alg. 1 and describe it below.

Norm concepts are *validated by humans* and *supported by data*. We create initial unnamed clusters of LLM norm descriptions. Humans inspect these clusters and create norm concepts by selecting 5-10 closely related examples of the concept. Humans also define a symbolic structure for the norm concept. Then, we perform k-NN augmentation of the concepts using untouched descriptions.

Then, humans inspect the newly augmented samples for each norm concept and mark 5-10 good and bad examples for each concept cluster. We re-perform k-NN augmentation for the untouched descriptions. Humans also create new norm concepts when they deem appropriate. We perform further iterations to discover more norm concepts. As the process progresses, the unnamed clusters

evolve and reveal new concepts. This iterative process helps us reliably amplify human mapping decisions to the entire dataset. We discovered 35 norm concepts during this phase with a coverage of 64% over $67k$ norm descriptions we collected. Annotation interface screenshots are presented in appendix I. We evaluate the results of the concept discovery process using human evaluation in §5. All annotators are CSS graduate students. We discuss this in detail in appendix C.

4.2.3 Symbolic Grounding

The HiL process creates norm concepts and maps them to conversation-specific descriptions. However, to fully ground the conversation in a norm concept, we must also align the conversation to the concept's symbolic structure. Hence, we leverage LLMs as *symbolic annotators* to identify instantiations of concept symbols in the conversation.

We provide the symbolic annotator LLM with a conversation, LLM-generated descriptions, and associated norm concept structure. We first ask to verify the *relevance* of the description to the conversation (filtering hallucinations) and the *correctness* of description-concept mapping (filtering incorrect HiL mappings). Then, we ask it to annotate violation status, actor roles, recipient roles, and violation details, if applicable. We provide the exact prompts and outputs in appendix E. We evaluate violation status annotations against human annotation in §5.

4.2.4 Automated Verification

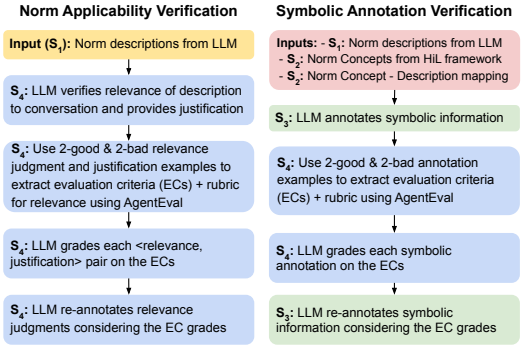


Figure 3: Automated Verification Flowcharts. S_* denotes pipeline stages from Fig. 2

As we use LLMs and semi-automated decision amplification, the dataset is prone to noise from hallucinations and incorrect mappings. Hence, we devise techniques to refine the dataset in a systematic manner. Fung et al. (2023) have shown that self-verification improves the LLMs output quality significantly. We aim to further improve this process by leveraging LLMs as multi-agent verifiers. To achieve this, we employ *AgentEval* tool proposed by (Arabzadeh and Clarke, 2024). We evaluate the initial, self-verified, and AgentEval-refined datasets against human judgments in §5. We refine both the LLM descriptions and symbolic grounding annotations using AgentEval.

Self-Verification is the technique of prompting an LLM to re-consider a judgment. This is shown to be an effective technique in improving LLM response quality (Weng et al., 2023; Fung et al., 2023). We employ this technique to refine descriptions, HiL mappings, and symbolic grounding of conversations. Exact prompts used for self-verification are in appendix D.

Multi-Agent Verification We further improve our refinement process via a principled and highly interpretable multi-agent verification framework. AgentEval (Arabzadeh et al., 2024) is a multi-agent framework for evaluating task utility. It takes the task description and 1-2 examples of successful and unsuccessful task runs as input. It generates a categorical rubric of criteria that helps determine the success/failure of future task runs.

The framework uses critic, quantifier, and verifier agents. The critic agent generates several criteria to evaluate the examples, the quantifier agent rates the examples on the criteria, and the verifier agent checks the robustness of each generated criterion. We add an evaluator agent on top of the criteria judgments to decide whether a data sample should be retained or not. We show the workflow charts in Fig. 3. We use instances of GPT-4o-mini for each agent in the framework. We provide the criteria generated for each evaluation workflow in appendix G.

4.3 Comparison with Existing Norm Datasets

Three of the prominent existing works that address social norms in conversation are NORMSAGE (Fung et al., 2023), NORMDIAL (Li et al., 2023) and NORMBANK (Ziems et al., 2023).

NORMSAGE presents a framework to obtain norms using LLMs and then perform self-verification. Their main contribution is the pipeline rather than a dataset. In contrast, we gather norm descriptions, operationalize a structured organization and filtering pipeline that efficiently uses and amplifies culturally proficient human annotation.

NORMDIAL is closest to our work as they propose a human-in-the-loop approach with LLM to obtain situation-specific norms. However, we in-

troduce *norm concepts* that aggregate situation-specific descriptions into more general behavioral concepts. We also present robust automated verification, which is evaluated using multi-point human evaluation. The scale of our dataset is larger.

NORMBANK uses situational attributes that align with the factual segment of our schema. However, unlike our work, these norms are not derived from real conversations but are instead defined by a list of high-level scenarios. Our dataset focuses on capturing nuanced human behavior within a specific culture, whereas NORMBANK aims for broader, more general cultural coverage.

5 Qualitative Evaluation

	Relevance		Mapping		Violations	
	qual	ret	qual	ret	qual	ret
LLMs	81	-	91	-	60.3	-
Self-Ver	82.2	73	93.4	85.6	64.3	74
MultiAgent	**88.4**	**91.3**	**94.8**	**93.8**	**66.1**	**81.4**

Table 3: Summary of human evaluation results of pipeline and refinement techniques. qual(ity) - % of correct samples in the refined data; ret(ention) - % of correct samples which passed refinement

Our cultural context dataset is prone to reliability issues from many sources, such as LLM hallucinations and scaling errors in the HiL process. Hence, we employ automated verification to rectify these errors. Now, we evaluate the quality of the obtained data and the refinement techniques via human evaluation. This allows us to quantify the reliability of the cultural context dataset as a resource. To that end, we obtain human judgments for several aspects of the dataset on a sampled subset. We observe that the automated verification-based filtering strategies significantly improve the dataset's quality. AgentEval's criteria-based refinement outperforms the self-verification strategy in all aspects.

Ideally, refinement techniques should retain *good* samples while discarding *bad* samples. Hence, we focus on two metrics for our evaluation: *quality* and *retention*. *Quality*, which is analogous to precision, is defined as the percentage of *good* samples in the post-refinement data. *Retention*, analogous to recall, is defined as the percentage of original *good* samples retained after refinement.

We randomly sample 726 description across 239 conversations for human evaluation annotation. Out of 726 examples, 580 are mapped to created norm concept structures. We focus our evaluation on three aspects: LLM hallucinations, symbolic annotation, and HiL scaling errors. Hence, we ask humans to (1) judge the *relevance* of the LLM-generated description to the conversation, (2) judge whether the mapping of the norm concept to the description is accurate, and (3) judge whether or not the norm concept is violated in that particular conversation. Each of the three decisions is a binary yes/no answer. We use three native culture annotators for these tasks. All annotators are graduate students who are CSS researchers as well.

We evaluate the initial LLM generations from §4.2.1, self-verification filtered, and AgentEval filtered datasets against human judgment data. We present the results in Tab. 3. We observe that the quality of the dataset improves upon self-verification in all three aspects. It further improves significantly upon multi-agent verification. We reduce hallucinations in norm description generations from 19% to 11.6%, improve description-norm concept mapping quality from 91% to 94.8%, and improve violation status quality from 60.3% to 66.1% using multi-agent verification refinement. We also note that multi-agent verification retains a significantly higher portion of data than self-verification. For norm descriptions, we retain 91.3% of correct descriptions (vs. 73% by self-verification); for norm concept mapping, we retain 93.8% (vs. 85.6%); and for violation judgments, we retain 81.4% (vs. 74%) data.

IAA: We also measure inter-annotator agreement to quantify the subjectivity of these tasks. For concept mapping, we obtain Krippendroff's alpha of 0.61, which points to moderate to strong agreement. For hallucinations, we obtain 0.74, and for violation status, we obtain 0.68. We also ask annotators to rate k-NN augmentations on how relevant they are to the assigned norm concept on a Likert scale of 1-5. This resulted in an average Likert score of 4.11. These results demonstrate high agreement and a high success rate of k-NN augmentation. We further present interesting qualitative visualizations of norm concepts in appendix A.

6 Downstream Task Evaluation

Our experiments aim to evaluate the *usefulness* of cultural information and graph-based schema. We use empirical performance on conversational understanding tasks as a proxy for *usefulness*. We focus on 2 classes of models: no-context models and cultural context models. Cultural context models

are further divided into 2 types: models that consume cultural context as *text* and as *a graph*. We use 3 models for our experiments: RoBERTa (Cui et al., 2021), LLama-3.1 (Dubey et al., 2024), and our ConvGraph model. We design a graph model that uses PLM as a node encoder and leverages the schema structure in Fig.1(a). We briefly discuss the datasets and tasks and then models used.

Datasets: We perform experiments on 3 tasks across 3 existing datasets. We conduct experiments with MPDD, CPED, and LDC CCU Chinese datasets (§3). We use emotion detection (all 3 datasets), sentiment detection (CPED), and dialogue act identification (LDC CCU, CPED) tasks to benchmark the models.

Tasks: For emotion detection, MPDD is annotated using 7 emotion labels, CPED with 13, and LDC CCU ZH with 9 labels. Dialogue act identification in CPED is a 19 class task, and the LDC CCU Chinese dataset uses 10 labels. CPED sentiment is a 3-way annotation. We use the original train/validation/test splits for CPED and MPDD datasets. For the LDC CCU Chinese dataset, we use the LDC2022E18 release as the train set, the LDC2023E01 release as the validation set, and the LDC2023E20 release as the test set. In this dataset, *neutral* label for emotion detection and *other* for dialogue act identification are over-represented. We also find that several samples from these classes are missed annotations. Hence, to avoid skewing the models, we down-sample these classes in all data splits to 1% of the actual labels. This makes these classes roughly the same size as the other classes.

6.1 Models

We experiment with 3 models: RoBERTa, LLama-3.1-8B-Instruct, & our ConvGraph model.

RoBERTa: We fine-tune the RoBERTa-Chinese-WWM-base (Cui et al., 2021) model on each task as sequence classification. We provide all the previous turns in the conversation and the current turn as inputs and predict class labels for respective tasks. We use cross-entropy loss to train the model. We use loss weighting to deal with class imbalance. Further details are provided in appendix B. We use the HuggingFace transformers library (Wolf et al., 2020) for all our experiments.

LlaMa-3.1-8B-Instruct: We perform QLoRA (Dettmers et al., 2024) fine-tuning of state-of-the-art LLM, Llama-3.1-8B-Instruct model using the same inputs as RoBERTa-Chinese-WWM model for the no-context experiments. Then, for cultural context experiments, we augment both factual component information and cultural information as text. We provide the exact input format in appendix F.

ConvGraph Model: For the no-context model, we create a graph from only conversational turns. The graph consists of one node with the conversation text. This node is connected to several child nodes with one turn each. We perform tasks as node classification. We use RoBERTa-Chinese-WWM-base to encode the dialogue and turn nodes. We use DGL (Wang et al., 2019) library to implement the graph model. We use two GraphSAGE (Hamilton et al., 2018) layers on top of the PLM encoders and then pass the node embedding to a final GraphSAGE layer for final task classification.

For the ConvGraph + cultural context model, we use the graph structure presented in Fig. 1(a). To encode cultural context nodes, which are in English, we use RoBERTa-base (Liu, 2019) as node encoder. We represent all edges as bi-directional edges. In this model, the contextual nodes such as norm concepts, relationships, settings, etc, are shared across conversations. This makes the entire data split (train/valid/test) into a single graph. We use randomly sample a neighborhood of 10 for each node during training and inference to make the computation tractable.

7 Results

Task/ Model	MPDD	CPED			LDC CCU	
	Em	Em	Sent	DA	Em	DA
	w-F1	w-F1	w-F1	w-F1	w-F1	w-F1
No Context Models						
RoBERTa	61.13	20.89	46.69	55.40	56.50	64.84
ConvGraph	61.81	21.42	47.66	56.28	54.83	65.62
Llama-3.1-8B-Instruct	45.48	24.73	53.91	54.15	60.11	67.15
Cultural Context Models						
Llama-3.1-8B-Instruct + Context	48.40	**29.81**	**55.00**	**60.16**	**62.16**	68.28
ConvGraph + Context	**64.34**	21.65	49.90	57.24	57.97	**71.74**

Table 4: Results on test sets. We report weighted F1 scores. Em-emotion, Sent-sentiment, DA-Dialogue Act.

We present the results on the test sets of each task of our experiments in Tab.4. We observe that the cultural context models outperform no-context models significantly in all the tasks. We report the frequency weighted-F1 score metric.

We mainly note the performance of cultural context models. Both Llama-3.1-8B and ConvGraph models perform significantly better when the cul-

tural context information is augmented. It is interesting to note that despite significant pre-training and post-training on extensive data, llama-3.1 models still benefit significantly from cultural context information. For the llama-3.1-8B model, performance on CPED emotion improves from 24.73 to 29.81, and CPED dialogue act performance improves from 54.15 to 60.16 F1.

It is also interesting to observe that the ConvGraph + cultural context model performs best on two tasks: MPDD emotions and LDC CCU dialogue acts. ConvGraph model only has $\approx$ 500M parameters as opposed to Llama-3.1-8B.

It is intriguing to note that the Llama-3.1-8B model significantly lags behind even the RoBERTa baseline on the MPDD emotion task. We posited that this might be due to a lack of robust multilingual capabilities. Hence, we trained the Llama-3.1-70b model on this task as well. But, it also reaches only 53.41 F1 on the task. We conclude that this might be due to the nature of the Llama class of models pre-training, post-training, and the annotation distribution for this particular dataset.

8 Conclusion

We propose a novel cultural context schema grounding pipeline. We introduce *Norm Concepts* which abstract over cultural beliefs. Using the pipeline and LLMs, we create high-quality cultural context data and perform human evaluation. Finally, we show that the dataset improves conversational understanding empirically.

Acknowledgments

We thank the reviewers for their insightful comments that helped improve the paper. This work was supported by NSF CAREER award IIS-2048001 and the DARPA CCU program. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, DARPA, or the US Government.

Limitations

We present a novel framework built upon LLMs and a human-in-the-loop approach. Both these approaches are prone to bias due to data and human components. Our evaluation is qualitative and hence relies on heuristic metrics. Our method requires expensive data collection and annotation protocol. We perform annotations for only one language. Cultural norm discovery using LLMs is limited due to the lack of depth on knowledge on some cultures for LLMs. Western bias due to training data might propagate into the dataset. This is a pioneering attempt to build large-scale datasets and hence requires careful usage of the data and technology. Our verification strategy also constitutes LLMs and hence that data still contains close to 10% noisy data event after validation.

LLM-generated cultural norms could potentially perpetuate stereotypes, contain hallucinations, bias, or other harmful distributional characteristics that could be challenging to detect and filter. In this work, we address these issues such as hallucinations, and stereotypes by (1) involving culturally proficient humans in the norm concept creation stage and (2) also performing principled automated filtering which is again evaluated against human annotation. In our pipeline, humans are not just aggregating the descriptions that they see, but rather they are using their cultural expertise to identify norm concepts that are prominent in their culture and are also supported by data. Despite these steps, the output is not necessarily completely absolved of these issues. On the other hand, human-generated datasets have also been shown to contain harmful biases (Hovy and Prabhumoye, 2021; Geva et al., 2019; Gautam and Srinath, 2024; Davidson et al., 2019; Das et al., 2024; Doughman and Khreich, 2022). LLMs are powerful NLP tools that allow us to explore tasks that were not practical before their advent. A case in point is our paper. It would have been highly impractical or hugely expensive to collect culturally proficient human annotations for cultural norms at this scale without LLMs. We believe that a pragmatic approach to this issue would be to carefully document and evaluate biases at each stage of the pipeline and invest community effort in building de-biasing techniques, bias-free training approaches, and so on. This could potentially lead us to creating fair(er) systems. We believe that a pragmatic approach of learning to build better guard rails, mitigating the harmful effects, and being aware of these biases could be a more fruitful path forward.

Ethics Statement

Cultural norm discovery using LLMs is limited due to the lack of depth on knowledge on some cultures for LLMs. Western bias due to training data might propagate into the dataset. This is a pioneering attempt to build large-scale datasets and hence re-

quires careful usage of the data and technology. Human annotation using cultural experts is a high-variance process as the experience of culture varies from region to region and person to person. We try to capture a high-level idea of culture in our work. It is possible that the data propagates biases in the society and hence requires careful usage. Our work is aimed to be a research artifact to help foster better models for cross-cultural interaction. This is by no means an end product to be used in large-scale applications.

References

Negar Arabzadeh and Charles Clarke. 2024. Fréchet distance for offline evaluation of information retrieval systems with sparse labels. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 420–431, St. Julian's, Malta. Association for Computational Linguistics.

Negar Arabzadeh, Siging Huo, Nikhil Mehta, Qinqyun Wu, Chi Wang, Ahmed Awadallah, Charles LA Clarke, and Julia Kiseleva. 2024. Assessing and verifying task utility in llm-powered applications. *arXiv preprint arXiv:2405.02178*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Muneera Bano, Didar Zowghi, and Jon Whittle. 2023. Exploring qualitative research using llms. *arXiv preprint arXiv:2306.13298*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Preprint*, arXiv:1812.01193.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 610–614, Marseille, France. European Language Resources Association.

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. *Preprint*, arXiv:2205.14727.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Amit Das, Zheng Zhang, Najib Hasan, Souvika Sarkar, Fatemeh Jamshidi, Tathagata Bhattacharya, Mostafa Rahgouy, Nilanjana Raychawdhary, Dongji Feng, Vinija Jain, et al. 2024. Investigating annotator bias in large language models for hate speech detection. *arXiv preprint arXiv:2406.11109*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jad Doughman and Wael Khreich. 2022. Gender bias in text: Labeled datasets and lexicons. *arXiv preprint arXiv:2201.08675*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Martha Finnemore. 1996. Norms, culture, and world politics: insights from sociology's institutionalism. *International organization*, 50(2):325–347.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Sanjana Gautam and Mukund Srinath. 2024. Blind spots and biases: Exploring the role of annotator cognitive biases in nlp. *arXiv preprint arXiv:2404.19071*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive representation learning on large graphs. *Preprint*, arXiv:1706.02216.

A Paul Hare. 2003. Roles, relationships, and groups in organizations: Some conclusions and recommendations. *Small group research*, 34(2):123–154.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Dell Hymes. 1972. Models of the interaction of language and social life. *Directions in sociolinguistics: The ethnography of communication/Holt, Rinehart & Winston*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023. Normdial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. *Preprint*, arXiv:2310.14563.

Ziming Li, Dookun Park, Julia Kiseleva, Young-Bum Kim, and Sungjin Lee. 2021. Deus: A data-driven approach to estimate user satisfaction in multi-turn dialogues. *arXiv preprint arXiv:2103.01287*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. 2023. Interactive concept learning for uncovering latent themes in large text collections. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5059–5080, Toronto, Canada. Association for Computational Linguistics.

Ben Prystawski, Michael Y. Li, and Noah D. Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. *Preprint*, arXiv:2304.03843.

Rajkumar Pujari, Chengfei Wu, and Dan Goldwasser. 2024. "we demand justice!": Towards social context grounding of political texts. *Preprint*, arXiv:2311.09106.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Kate Roberts, Anthony Dowell, and Jing-Bao Nie. 2019. Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC medical research methodology*, 19:1–8.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of us politicians through the lens of moral foundation theory. In *Proceedings of the ninth international workshop on natural language processing for social media*, pages 1–13.

Roger C Schank and Robert P Abelson. 1977. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press.

Muzafer Sherif. 1936. The psychology of social norms.

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *23rd International Conference on Intelligent User Interfaces*, pages 293–304.

Aarohi Srivastava, Abhinav Rastogi, and Abhishek Rao et. al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Harry Charalambos Triandis et al. 1994. *Culture and social behavior*. McGraw-Hill New York.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners

with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

A Norm Concept Visualization

Further, we present a qualitative visualization of norm concept distribution over conversational settings in Fig.4. We present the visualizations for 1) Norm of Formal Address, 2) Norm of Children obeying Parents' wishes and 3) Respecting the Doctor's expertise. We observe that norms are prominent in different fields. Formal address norm is highly prominent in a company setting, while children obeying parents' wishes is more prevalent in family settings. This is in accordance with the general expectations. This shows that the norm concepts capture interesting aspects of conversational context. Norm (3) is highly prominent in family and hospital settings.

It is interesting to note that we only have field values at a conversational level, and hence, there could be multiple fields with the same conversation. The conversations are not segmented neatly on a scene-to-scene basis, especially in the LDC CCU dataset. This explains the noise distribution of certain norms. It is also noteworthy that the norm descriptions are generated based on conversational content. The field of the conversation alone doesn't always determine the conversational content.

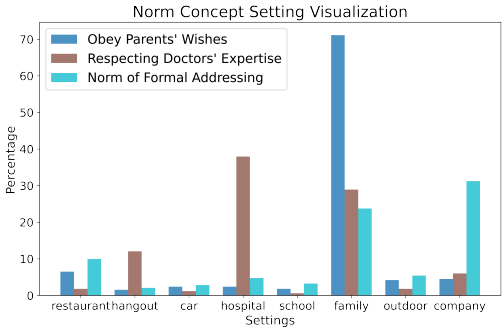


Figure 4: Comparison of Conversation Field Distribution Across Various Norm Concepts

B Reproducibility

For our downstream task experiments, we use a community cluster with several 80G and 40G A100 GPUs. We use a learning rate of 1e-4 for llama QLoRa training. We use r=16 with 8-bit quantization. For the graph model, we use the DGL library with distributed training using several A100 GPUs. Our llama training took 12 hours on average for each task. We ran several development iterations before arriving at the final prompt structure. The graph model training took 4 hours on aver-

age for each task, when augmented with cultural schema information. RoBERTa models take 10-30 minutes to converge in training depending on the size of the dataset. We will release all our code and datasets under MIT license upon acceptance. We extensively use ChatGPT, Claude, Github co-pilot for our coding requirements. We paid ∼400 USD for using OpenAI API for data collection and experiments. We use Grammarly to aid in draft improvement.

C Annotation Guidelines

All annotators used for both HiL concept discovery and human evaluation are graduate students, 3 of whom are also CSS researchers. We have a quid-pro-quo data annotation system in place in our lab. We contribute hours of annotations to other projects and we can get similar hours back when we need them. We use 3 native Chinese annotators, who were born and raised in mainland China for our annotations. We also use 2 non-Chinese annotators. They help in book-keeping tasks in the norm concept generation. They moderate and organize the interactive process sessions which are collaborative. The HiL process described in §4.2.2 is a collaborative annotation process. All the annotators discuss and reach a consensus before adding a norm concept to the database. In contrast, human evaluation annotation described in §5 is an individual annotation task. Hence, we report inter-annotator agreement for these judgments. This annotation is performed by native culture annotators.

D LLM Prompts and Generations

Prompts used to extract Conversation-Specific Social Norms and Violations from GPT-3.5

1. Translate this conversation into English.
2. List the people mentioned in the conversation and the social relationships between them.
3. List the Chinese cultural norms applicable to this situation. Are there any cultural norm violations observed in this situation? If yes, list them. List the observed and potential effects by index for each violation.
4. Summarize the conversation in 3-4 sentences.

Input

Translated Conversation:
Mrs. Zuo: What is that foolish girl worth giving anything to? Zho Zpeng is such a person!
Mr. Zuo: Oh, wife, why do you always say things that are not conducive to unity? It's reasonable for him to go and give something to his classmate!
Zho Zpeng: Dad, Mom, I'm back!
Mrs. Zuo: I'm afraid you've been drugged and your soul has been seduced by that foolish classmate of yours!
Mr. Zuo: Wife, do you sound like a mother when you speak like that? Why do you always oppose their relationship?
Zho Zpeng: Dad, Mom, please stop talking. I have my own opinions on this matter. I know what I should do and what I shouldn't do. I will never act recklessly. Please trust me!

LLM Description Generations

Summary:
The conversation revolves around Zho Zpeng, his parents, and his relationship with a classmate. Mrs. Zuo disapproves of the classmate and questions the worth of giving her anything, while Mr. Zuo defends his son's actions. Zho Zpeng expresses his independence and asserts that he knows what he should and shouldn't do, asking his parents to trust him.

Norms:
Respect for parents: Filial piety and respect for parents are highly valued in Chinese culture. Children are expected to listen to and obey their parents' opinions and decisions.
Unity within the family: Maintaining harmony and unity within the family is important in Chinese culture. Conflicts or disagreements that disrupt family unity are generally discouraged.
Social relationships and obligations: Chinese culture places importance on maintaining positive relationships and fulfilling social obligations, such as giving gifts to classmates or friends.

Violations:
Disrespectful language: Mrs. Zuo uses derogatory language to refer to Zho Zpeng's classmate, which shows a lack of respect for others.
Opposition towards Zho Zpeng's relationship: Mrs. Zuo continuously opposes the relationship between Zho Zpeng and his classmate, going against the cultural norm of respecting children's autonomy and choices in relationships.

Effects:
Disrespectful language - Observed effect: It can create tension and animosity between Mrs. Zuo and Zho Zpeng.
Opposition towards Zho Zpeng's relationship - Observed effect: It causes disagreement and arguments between Mrs. Zuo and Mr. Zuo, highlighting a lack of unity within the family.

E Symbolic Annotation Analysis

Symbolic Annotation & Verification Prompts
As a two-step cultural and social norms analysis assistant, your task is to evaluate a provided conversation from the context of Chinese culture based on a given social norm and a corresponding norm concept. Your analysis should be comprehensive, considering factors such as age, relationships, settings (e.g., work, family, friends), and the topic of the conversation.

Steps for Analysis:

1. **Evaluate Social Norm and Norm Concept Compatibility** - **Task:** Judge whether the provided social norm aligns with the given norm concept. - **Action:** State if the social norm **matches** or **doesn't match** the norm concept. - **Justification:** Provide a concise reason for your judgment.

2. **Assess Relevance to the Conversation** - **Task:** ..

3. **Determine Social Norm Violation** - **Task:** Judge whether the social norm was **adhered to** or **violated** in the conversation. - **Justification:** Provide a concise reason for your judgment.

4. **Annotate Conversation-Specific Details**

- **Enactor Role:** ...
- **Acceptor Role:** ...

5. **Violation Analysis** *(Only if a violation occurred)* If the norm was violated, provide the following additional details: - **Violating Action:** A brief description of the action that caused the violation (e.g., "badmouthing parents").

- **Violator Role:** ..
- **Victim Role:** ..
- **Violator Emotion:** ..
- **Victim Emotion:** ..

Response Format: Your response must adhere to the format below:

Social Norm - Norm Concept Compatibility: <match/doesn't match> Compatibility Justification: <short justification>

Only if compatible Relevance: <relevant/irrelevant> Relevance Justification: <short justification>

Enactor Role: <sitatuion specific social role, strictly not the name, of the person> Acceptor Role: <sitatuion specific social role, strictly not the name, of the person>

Violation Status: <adhere/violate> Violation Status Justification: <short justification>

Only if a violation occurs
Violating Action: <short phrase>
Violator Role: <sitatuion specific social role, strictly not the name, of the person>
Victim Role: <sitatuion specific social role, strictly not the name, of the person>
Violator Emotion: <one of 9 basic emotions>
Victim Emotion: <one of 9 basic emotions>

Input:

Conversation from Chinese culture:
Zuo Zhengpeng: Lihua, how much longer until the baby is due?
Xu Lihua: Probably next week!
Zuo Zhengpeng: Oh, the baby is moving!
Xu Lihua: Look at you, how embarrassing!
Xu Lihua: What's the rush? We still have a few days. We'll talk about it then. You're eager to become a father, aren't you?
Zuo Zhengpeng: No! By then, it might be too late. It's better to go to the hospital a few days early. Transportation here is inconvenient. If anything happens, we could be in big trouble.

Social Norm:
Respect for elders: It is common in Chinese culture to show respect to older family members, such as parents and grandparents.

Norm Concept Name: Respect for family elders
Norm Concept Description: Respecting the wisdom, experience, and authority of elder members in the family hierarchy.
Norm Concept Potential Violation Sketch: Showing disrespect and ignoring suggestions and advice of any elder members of the family.
Norm Concept Scenario: family
Enactor Role: any younger family member
Acceptor Role: elder family members such as parents, uncle, grandparents, etc.

GPT Output:

Norm-Concept Compatibility: match

Relevance: relevant
Relevance Justification: The conversation takes place within a family context, discussing the impending birth of a child, which inherently involves respect for family roles and responsibilities.

Enactor Role: younger family member
Acceptor Role: elder family member

Quality Judgment: accurate
Justification: The annotations correctly identify the compatibility of the social norm with the norm concept, the relevance to the family context, and the adherence to the norm in the conversation, reflecting a respectful interaction between the younger and elder family members.

F Downstream Task Setup

Instruction and final <task-name> prompt are not provided for the RoBERTa model.

LlaMa-3.1-8B-Instruct, RoBERTa Models Input:

Instruction: You are a helpful assistant who predicts the <task-name> of the conversational turn in a Chinese conversation. You are given the cultural context surrounding the conversation, the prior conversation, and the current turn. Predict the <task-name> of the current turn. Choose one of <label set>.

Prior Conversation:
<turn 1>
<turn 2>

Current Turn:
<turn-n>

<task-name>:

LlaMa-3.1-8B-Instruct + Cultural Context Model Input:

```
Instruction: You are a helpful assistant who predicts
the <task-name> of the conversational turn in a Chinese
conversation.  You are given the cultural context
surrounding the conversation, the prior conversation,
and the current turn. Predict the <task-name> of the
current turn. Choose one of <label set>.

Prior Conversation:
<turn 1>
<turn 2>

Conversation Settings:
Synopsis: ..
Speaker Count: ..
Speaker Sex: ..
...

Relevant Cultural Information:
1) **Norm Concept**:
Theme: <theme-name>
Description: ...
Settings: ..
Violation Sketch: ...

Specific Norm:
<norm-description>

2) **Norm Concept**:
...

**Potential Violations**:
1) <violation-description-1>
..

Current Turn:
<turn-n>
*<task-name>**:
```

G AgentEval Task Criteria

H Data Sources and Statistics

I Annotation GUI

J Norm Concept Visualization

K Schema Example

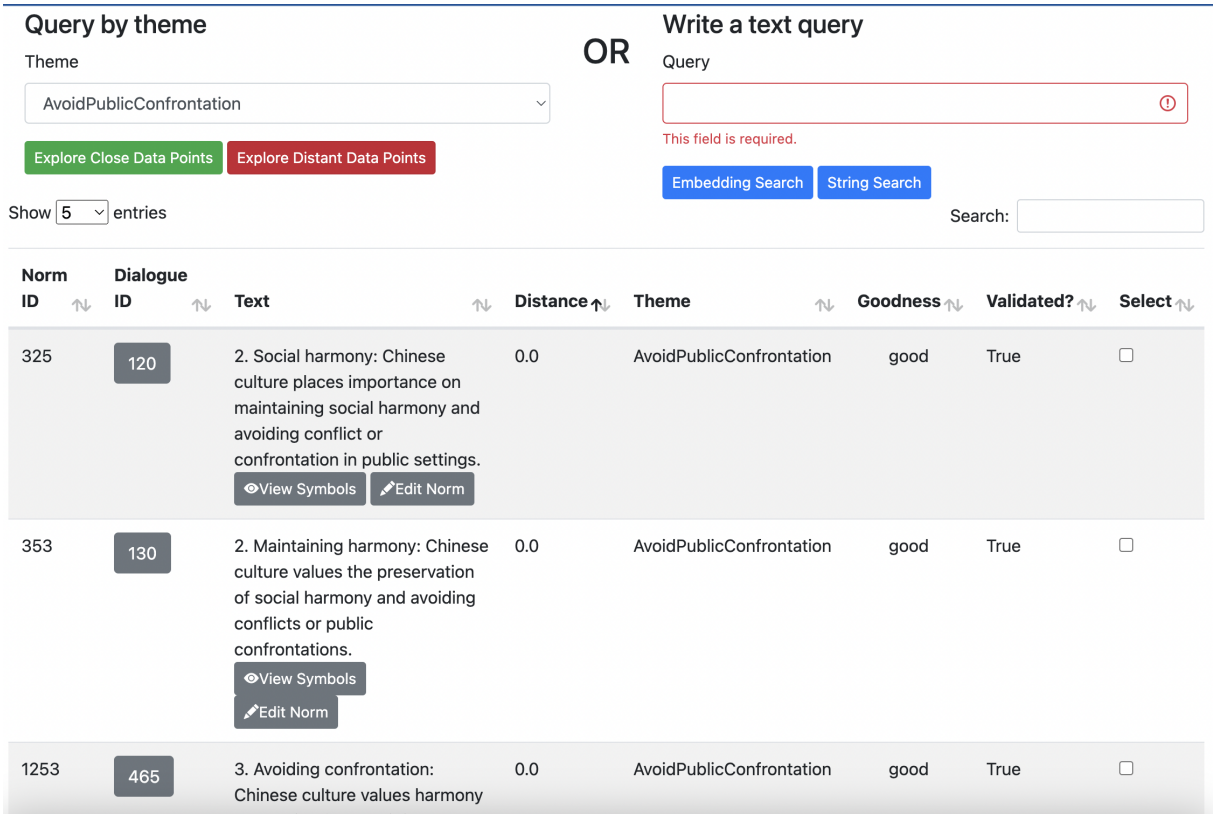


Figure 5: Annotation Interface for Norm Concept Discovery

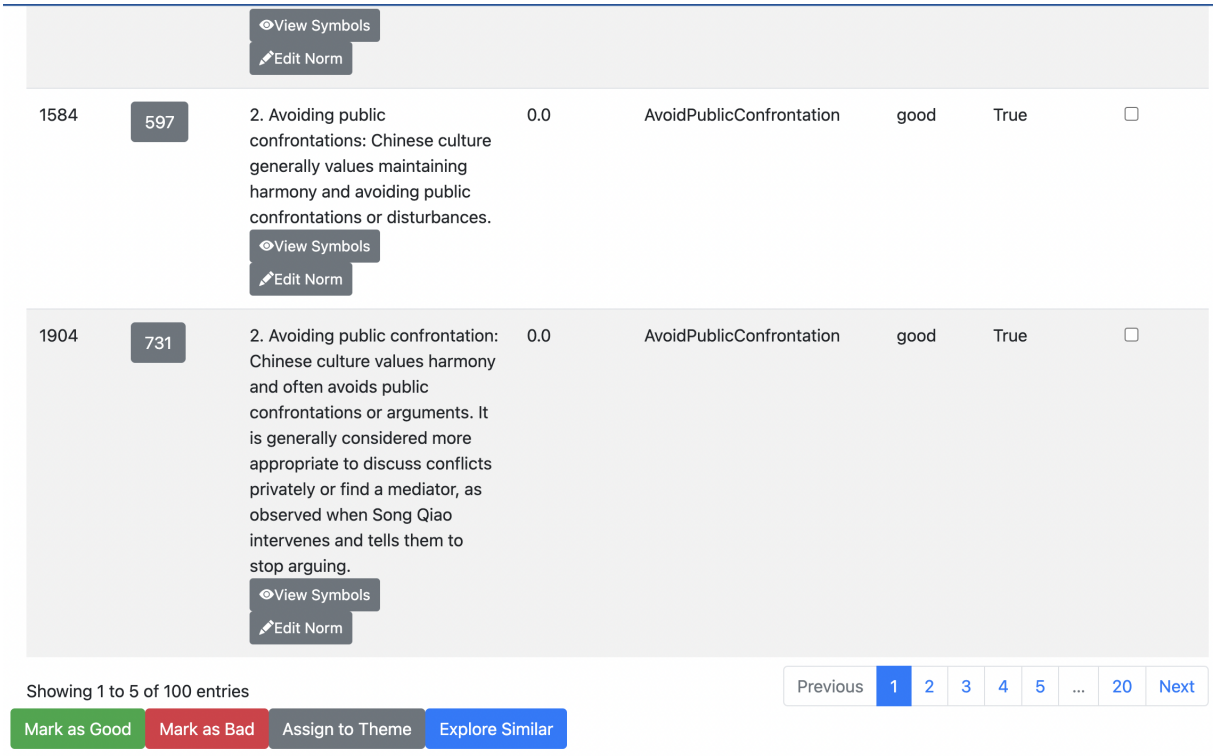


Figure 6: Annotation Interface for Norm Concept Discovery

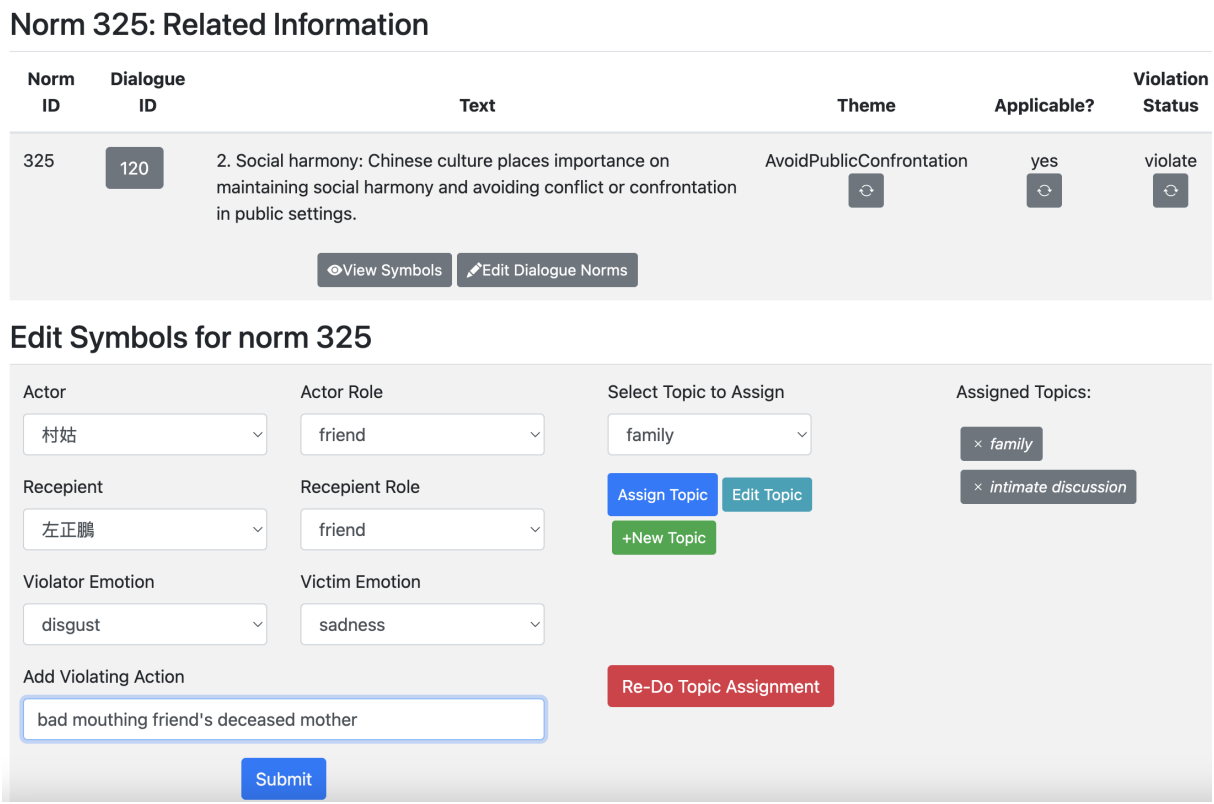


Figure 7: Human Validation and Symbolic Annotation of Cultural Context Annotation Framework

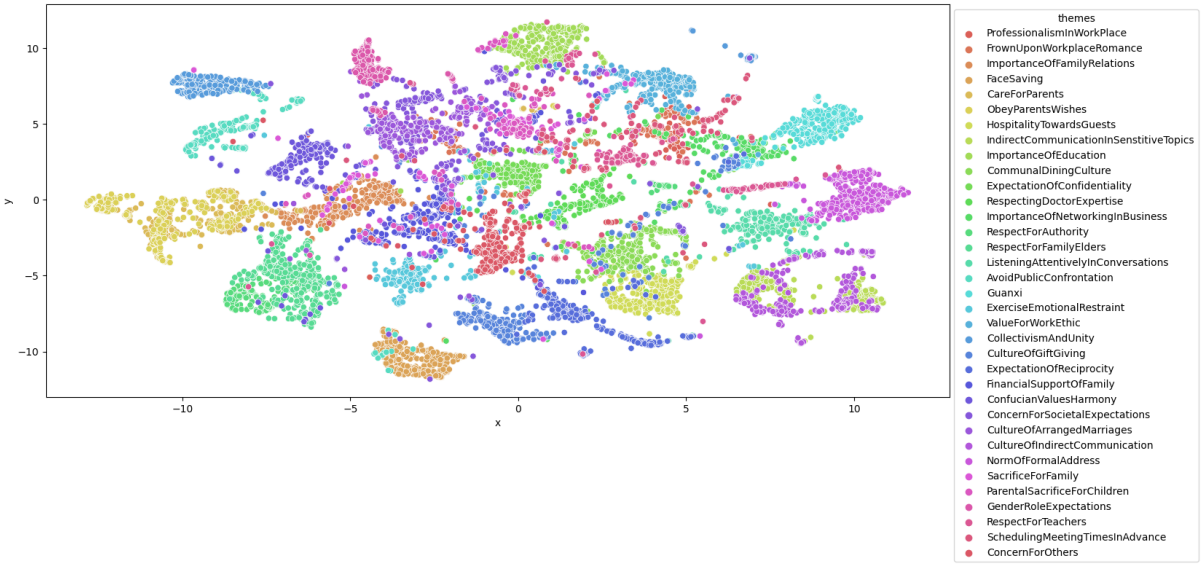


Figure 8: Norm Concept Visualization

Criteria	Description	Accepted Values
Task: Relevance Judgment		
Clarity	The degree to which the reasoning behind the relevance judgment is clear and understandable.	1 - very unclear, 2 - unclear, 3 - neutral, 4 - clear, 5 - very clear
Contextuality	The extent to which the judgment considers the specific context of the conversation including relationships, settings, and cultural nuances.	1 - not contextualized, 2 - poorly contextualized, 3 - moderately contextualized, 4 - well contextualized, 5 - very well contextualized
Appropriateness	How well the social norm applies to the situation discussed in the conversation.	1 - not appropriate, 2 - minimally appropriate, 3 - somewhat appropriate, 4 - appropriate, 5 - highly appropriate
Cultural Relevance	The degree to which the social norm reflects significant aspects of Chinese culture relevant to the conversation.	1 - not culturally relevant, 2 - minimally relevant, 3 - somewhat relevant, 4 - culturally relevant, 5 - highly culturally relevant
Consistency	The consistency of the judgment with established norms and expectations in Chinese cultural contexts.	1 - highly inconsistent, 2 - inconsistent, 3 - neutral, 4 - consistent, 5 - highly consistent
Relationship Dynamics	The consideration of the relationships between the individuals involved in the conversation and how that influences the relevance of the social norm.	1 - not considered, 2 - weakly considered, 3 - moderately considered, 4 - strongly considered, 5 - very strongly considered
Setting Analysis	Evaluation of how the setting of the conversation (e.g., family, workplace, casual gathering) impacts the relevance of the social norm.	1 - no impact, 2 - minor impact, 3 - moderate impact, 4 - significant impact, 5 - critical impact
Norm Visibility	The extent to which the social norm is visible or recognized in the context of the conversation.	1 - not visible, 2 - slightly visible, 3 - moderately visible, 4 - visible, 5 - very visible
Social Hierarchy	How well the judgment incorporates aspects of social hierarchy or status within the Chinese cultural context.	1 - not addressed, 2 - poorly addressed, 3 - addressed, 4 - well addressed, 5 - very well addressed
Timeliness	Consider how the relevance of the social norm may change over time or in different historical contexts.	1 - not timely, 2 - slightly timely, 3 - moderately timely, 4 - timely, 5 - very timely
Evidence	The amount and quality of evidence or examples provided to support the relevance judgment.	1 - no evidence, 2 - minimal evidence, 3 - some evidence, 4 - good evidence, 5 - strong evidence
Norm Specificity	The degree to which the social norm discussed is specific and detailed rather than vague.	1 - very vague, 2 - vague, 3 - somewhat specific, 4 - specific, 5 - very specific
Emotional Tone	Evaluation of how the emotional tone of the conversation affects the relevance of the social norm.	1 - negative impact, 2 - slight negative impact, 3 - neutral impact, 4 - slight positive impact, 5 - positive impact
Discourse Style	The impact of the discourse style (formal, informal, persuasive, etc.) on the relevance of the social norm.	1 - no impact, 2 - minimal impact, 3 - moderate impact, 4 - significant impact, 5 - critical impact
Dissonance Level	The level of dissonance between the social norm and the expressed opinions or behaviors in the conversation.	1 - highly dissonant, 2 - moderately dissonant, 3 - neutral, 4 - somewhat aligned, 5 - highly aligned
Cultural Evolution	Consideration of how modern trends or changes in society may affect the relevance of traditional social norms.	1 - outdated, 2 - slightly outdated, 3 - somewhat relevant, 4 - relevant, 5 - highly relevant
Collective Perspective	How well the judgment considers the views of the collective society versus individual perspectives.	1 - individual-focused, 2 - slightly collective, 3 - moderately collective, 4 - largely collective, 5 - entirely collective
Norm Adoption	Recognition of how widely the social norm is adopted or practiced within Chinese society.	1 - not adopted, 2 - rarely adopted, 3 - somewhat adopted, 4 - widely adopted, 5 - universally adopted
Age Sensitivity	The extent to which the relevance judgment is sensitive to the age differences of the individuals involved.	1 - not sensitive, 2 - slightly sensitive, 3 - moderately sensitive, 4 - sensitive, 5 - very sensitive
Analytic Depth	The depth of analysis provided in the reasoning for the relevance judgment.	1 - very superficial, 2 - superficial, 3 - moderate depth, 4 - deep, 5 - very deep
Feedback Responsiveness	The degree to which the judgment accounts for feedback or reactions from the individuals in the conversation.	1 - not responsive, 2 - slightly responsive, 3 - moderately responsive, 4 - responsive, 5 - very responsive
Rule Conformity	How closely the judgment conforms to established social rules and expectations within the context.	1 - highly non-conformant, 2 - non-conformant, 3 - somewhat conformant, 4 - conformant, 5 - highly conformant

Table 5: Evaluation Criteria Generated by AgentEval for *Relevance* of *Norm Description* to *Conversation* Judgment

Criteria	Description	Accepted Values
Task: Norm Interpretation and Evaluation		
Contextual Influence Evaluation	Evaluate how the context (e.g., setting, relationships) influences the interpretation of the norm.	high influence, medium influence, low influence, no influence
Cultural Appropriateness Assessment	Judge whether the norm is appropriate within the cultural context of the conversation.	appropriate, inappropriate
Generational Perspective Evaluation	Assess how generational differences impact the conversation's norms.	significant difference, some difference, no difference
Emotional Weight Assessment	Determine the emotional significance of the norm within the conversation.	high weight, medium weight, low weight
Implications of Norm Violation	Assess the potential consequences of violating the social norm.	severe consequences, moderate consequences, mild consequences, no consequences
Responsibility Assessment	Evaluate who holds the primary responsibility for adhering to the social norm in the conversation.	enactor, acceptor, both
Social Norm Awareness Evaluation	Judge the awareness of the social norm by the involved parties.	fully aware, somewhat aware, not aware
Feedback Necessity Assessment	Determine if feedback is necessary based on the adherence or violation of the norm.	necessary, not necessary
Coping Strategy Evaluation	Evaluate the coping strategies employed by individuals in response to a norm violation.	effective, somewhat effective, ineffective
Communication Style Analysis	Analyze the communication styles used in relation to the social norm.	formal, informal, assertive, passive, aggressive

Table 6: Evaluation Criteria Generated by AgentEval for *Symbolic Annotation Quality* Judgment

Dataset	Turn-Level				Conversation-Level									Corpus	
	Em.	DA	Sent.	Sp.-List. Reln.	Summ.	Norms			Violations			Effects		Field	Norm Concepts	
						Desc	Valid	Symb. Attr	Desc	Valid	Symb. Attr	Desc	Valid		Conc.	Assg. Norms
MPDD	A	LM	-	A	GPT	10,637 (GPT)	900 (H)	726 (H)	5,721 (GPT)	213+65915 (H+GPT)	66,128 (H+GPT)	14,521 (GPT)	213 (H)	4,141 (A)	35 (H)	422 (H)
CPED	A	A	A	LM	GPT	32,209 (GPT)			17,865 (GPT)			25,866 (GPT)		11,832 (A)		
LDC CCU	A	A	-	LM	GPT	23,306 (GPT)			15,605 (GPT)			14,908 (GPT)		7,554 (A, LM)		36,954 (k-NN)

Table 7: Sources and Counts of Collected Schema Grounding Dataset. A - gold annotation; LM - Llama-70b generated; GPT - GPT-3.5 generated; H - human annotation; kNN - interactive k-nearest neighbors; Em - emotion; DA - dialogue act; Sent - sentiment; Sp_List Reln - speaker_listener relationship; Desc - descriptions; Symb_Attr - symbolic attributes; Valid - Validated; Conc - concepts; Assg_Norms - norms assigned to concepts;

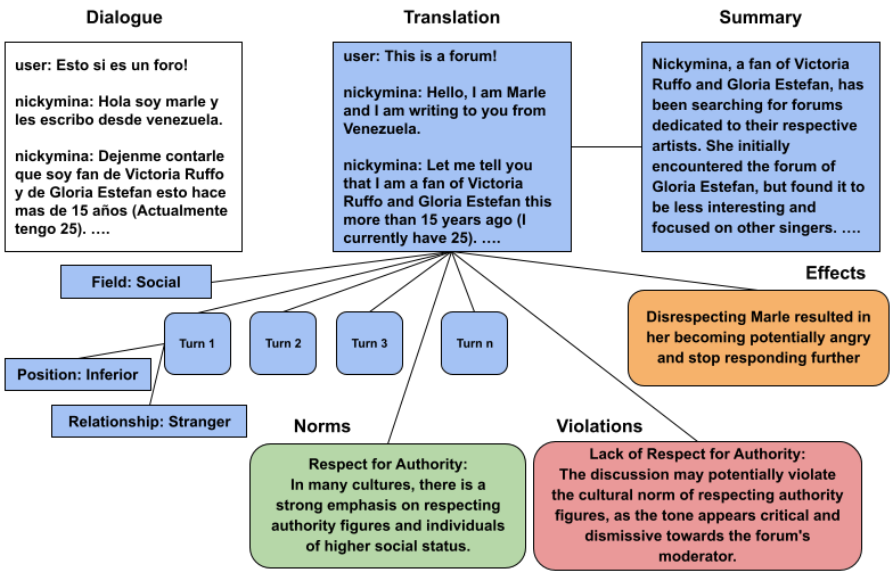


Figure 9: An Example Instance of Schema Augmented Conversation