Multi-Cultural Norm Base: Frame-based Norm Discovery in Multi-Cultural Settings

Anonymous ACL submission

Abstract

Sociocultural norms serve as guiding principles for personal conduct in social interactions within a particular society or culture. The study of norm discovery has seen significant development over the last few years, with various interesting approaches. However, it is difficult to adopt these approaches to discover norms in a new culture, as they rely either on human annotations or real-world dialogue contents. This paper presents a robust automatic norm discovery pipeline, which utilizes the cultural knowledge of GPT-3.5 Turbo (ChatGPT) along with several social factors. By using these social factors and ChatGPT, our pipeline avoids the use of human dialogues that tend to be limited to specific scenarios, as well as the use of human annotations that make it difficult and costly to enlarge the dataset. The resulting database - Multi-cultural Norm Base (MNB) - covers 6 distinct cultures, with over 150k sociocultural norm statements in total. A state-of-the-art Large Language Model (LLM), Llama 2, fine-tuned with our proposed dataset, shows remarkable results on various downstream tasks, outperforming models fine-tuned on other datasets significantly.

1 Introduction

Sociocultural norms are informal rules or guidelines that dictate acceptable behavior within a particular society or culture (Morris et al., 2015). These norms encompass a wide range of behaviors, including manners, customs, values, and traditions. They govern how individuals interact with one another and shape societal expectations regarding appropriate conduct in various contexts. With the rapid development of AI in the last decade, it is crucial to define effective methods for discovering and assessing the cultural knowledge of AI systems, especially the knowledge of sociocultural norms.

The study of cultural norm discovery has witnessed significant development in recent years. SOCIAL-CHEM-101 (Forbes et al., 2020), one of the earliest corpora, introduces social norms represented in a Rule of Thumb (RoT) format. NormBank (Ziems et al., 2023) is another large-scale corpus of norms that contains situational norms within a multivalent sociocultural frame. While these datasets have high-quality samples and can be applied to many culture-related tasks, they are constructed by humans, which is very time-consuming and costly. In response to this problem, Fung et al. (2023) introduced NormSage, a norm dataset constructed with a fully automated pipeline. Norm statements in NormSage are extracted by prompting Large Language Models (LLMs) with dialogue-based contents. The norms are then fed to a self-verification process to ensure their quality. While NormSage showcases a promising direction for automatic norm discovery, it is based on real dialogue data, which may not be available in different cultures and can be limited to specific domains. Moreover, social norms, relevant to specific frames, should possess the flexibility to be applicable across diverse dialogues, instead of being bound to a single specific conversation.

To address the above challenges, in this paper, we present an automated frame-based pipeline for norm dataset construction using ChatGPT in a multi-cultural setting. Socio-cultural norms are often strongly associated with several social factors (Zhan et al., 2023), and we refer to the combination of social factors as situational frames. Norms in the proposed dataset are generated by prompting ChatGPT with situational frames as the context, instead of using real-world dialogue content like existing works. These frames consist of carefully chosen social factors (culture, social relation, power distance, and so on) which help to align the norm generation process. In this way, we will not have to collect dialogue data for specific cultures and can easily expand the dataset. Once the norms are extracted, we evaluate them both intrinsically and extrinsically. For the former, we use human evaluation to assess the quality of the extracted norm statements. For the latter, we employ the constructed norm database in various downstream tasks to prove the adaptability as well as the performance of our proposed dataset. To summarise, our contributions are as follows:

- We propose an automatic pipeline for extracting socio-cultural norm statements in multiple cultures. This pipeline makes use of the implicit cultural knowledge of ChatGPT, as well as a set of carefully chosen social factors, to derive meaningful norm statements. In this way, we address the aforementioned problems of pioneering works. By using social factors and ChatGPT, we avoid the high costs of human annotation. Additionally, our social factors can also replace human dialogues, which tend to be limited to specific domains (Fung et al., 2023).
- With the proposed pipeline, we construct the Multi-Cultural Norm Base (MNB) dataset and make it publicly available to the research community. The dataset contains 150k sociocultural norm statements for 6 different cultural backgrounds, extracted from 29k situational frames. MNB is also one of the very few datasets that feature multi-cultural settings. We will make the dataset and code publicly available upon paper publication.
- We conduct extensive experiments to analyze the quality of MNB, as well as to demonstrate the benefits of MNB in various downstream tasks. Intrinsic evaluation results highlight both the strengths and weaknesses of our method. We observe that using ChatGPT for norm extraction results in correct and insightful norms. At the same time, the model cannot utilize all of the given social factors, which, in many cases, leads to norms being too general. On the other hand, however, extrinsic experimental results show that MNB can generalize well across multiple related datasets and their corresponding benchmarks, outperforming other datasets significantly.

2 Related Work

2.1 Commonsense Knowledge Bases

Commonsense Knowledge Bases (CKBs) encapsulate essential information that mirrors human everyday understanding and reasoning, covering broad aspects such as relational taxonomies (Liu and Singh, 2004), logical associations (Zhang et al., 2018; Elsahar et al., 2018), and the underlying principles of causality and mechanics (Talmor et al., 2019; Bisk et al., 2020). Following Cyc's establishment (Lenat, 1995), there has been a significant advancement in the development of expansive, human-curated CKBs (Liu and Singh, 2004; Speer et al., 2017; Forbes et al., 2020; Bisk et al., 2020; Hwang et al., 2021; Mostafazadeh et al., 2020; Ilievski et al., 2021). Notably, ConceptNet (Speer et al., 2017) exemplifies a comprehensive commonsense knowledge graph, characterized by its structured representation of knowledge in entity-relation-entity triples. The ATOMIC (Sap et al., 2019) advances this domain by cataloging social interaction dynamics through nearly 880,000 annotated triples. Its enhanced iteration, ATOMIC2020 (Hwang et al., 2021), further integrates ConceptNet's relational framework with additional novel relations, thereby constructing a more elaborate CKB focused on event-related dynamics. Moreover, GLUCOSE (Mostafazadeh et al., 2020), derived from narrative texts in ROC Stories (Schwartz et al., 2017), delineates a framework for understanding causal relationships and effects based on foundational events, presenting a nuanced exploration of commonsense dimensions.

2.2 Sociocultural NormBase Construction

SOCIAL-CHEM-101 (Forbes et al., 2020) introduced a comprehensive dataset of social and moral guidelines, established through a crowdsourcing approach to gathering descriptive norms from various situations using rules-of-thumb as fundamental elements. Another critical contribution is from (Ziems et al., 2023), who devised a layered classification system for social constraints, dubbed the Situational Constraints for Social Expectations, Norms, and Etiquette (SCENE), and subsequently recruited participants to label the extensive SCENE categories. Our methodology diverges significantly from that of NormBank by implementing an automated system to discover sociocultural norms, in contrast to the reliance of NormBank on manual annotation. Moreover, the research by (Fung et al., 2023) introduced the NormSage framework, aimed at identifying norms embedded within conversations, utilizing LLM prompting and self-verification techniques, and drawing from real-life scenarios like negotiations, casual discussions,

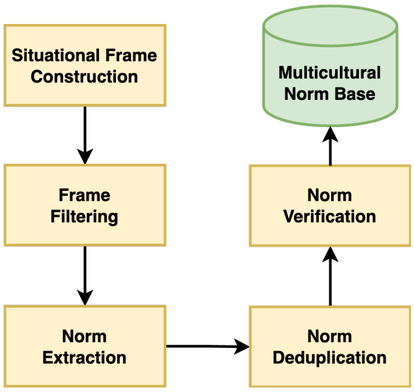

Figure 1: Proposed norm discovery pipeline.

and documentaries. Our approach sets itself apart from NormSage by focusing on extracting norms through a social-cultural lens exclusively, omitting dialogue-based information.

3 Building Multi-cultural Norm Base

In this section, we describe our proposed automatic pipeline for collecting socio-cultural norms for various cultures. The following subsections will discuss the overall pipeline, as well as provide a detailed explanation for each step in the pipeline. For simplicity, the term socio-cultural norm will be referred to as norm or social norm for short.

3.1 Overall Pipeline

The overall norm discovery pipeline is illustrated in Figure 1. Starting from a collection of situation frames, we begin by filtering invalid frames, followed by performing norm extraction, deduplication, and verification to construct the multicultural norm base.

3.2 Situational Frame Construction

Social norms are context-specific patterns that govern behavior in a given situation (Morris et al., 2015). Therefore, we design situational frames to ground meaningful norms and create diversity in the proposed dataset. Following the works of social factor taxonomy (Hovy and Yang, 2021) and SocialDial (Zhan et al., 2023), these situational frames consist of several social factors that mimic the conversations between two speakers. Specifically, there are 10 key social factors in a frame, and these factors are categorized as either conversation-related factors (*Norm Category, Conversation Topic, Conversation Location, Culture, Formality*) or speaker-related factors (*Age, Gender, Social Relation, Social Distance, Power Distance*). Each of these social factors can take a range of values, some of which are sourced from SocialDial and LDC (Li et al., 2022).

Conversation-related Factors. In each situational frame, *Norm Category* can take values from greetings, requests, apologies, persuasion, and criticism. *Formality* is characterized as either formal or informal. *Conversation Location* spans various settings, including open areas, online platforms, homes, police stations, restaurants, stores, and hotels. *Conversation Topic* covers a wide array of subjects, such as sales, everyday life trivialities, office affairs, school life, culinary topics, farming, poverty assistance, police corruption, counterterrorism, and cases of child disappearance. *Culture* refers to the cultural background of a conversation, which can be derived from one of the following values: American, British, Canadian, Indian, Afghan, and Chinese. We selected the American and Canadian to represent North American cultures. We chose the British as the representative of European culture. For Asian cultures, we identified Afghan as representing Middle Eastern culture and considered Chinese and Indian due to their status as the most populous countries in the world.

Speaker-related Factors. Regarding the speaker-related factors, *Social Distance* encompasses five distinct values: family, friends, romantic partners, working relationships, and strangers. *Social Relation* covers the following cases: peer-to-peer, elder-junior, chief-subordinate, mentor-mentee, student-professor, customer-server, and partner-partner. *Age* describe the age group of each speaker in the conversation, which can take the following values: child, teenager, adult, middle-aged adult, senior adult, and elderly. Similarly, *Gender* represents the gender of each speaker, which is categorized as either male or female. Lastly, *Power distance* is the perceived degree of inequality between the two speakers. This factor can take values from lower, equal, or higher, which indicates the inequality of the first speaker with respect to the second speaker.

3.3 Frame Filtering

With the values of each social factor predefined in the previous section, we then proceed to remove invalid situational frames. Invalid frames are those considered to have combinations of values that hardly represent real-world scenarios (eg. "a student and a professor discussing life trivialities in a police station", or "two colleagues discussing

Header: Consider the following context of a conversation between two speakers:

- The topic of the conversation is **{Conversation Topic}**.
- The sociocultural norm category is **{Norm Category}**.
- The cultural background is **{Culture}**.
- The level of formality is set to **{Formality}**.
- The location is **{Conversation Location}**.
- The social relation between the speakers is **{Social Relation}**.
- The gender of Speaker 1 is **{Speaker 1's Gender}**.
- The gender of Speaker 2 is **{Speaker 2's Gender}**.
- Speaker 1 is a/an **{Speaker 1's Age Group}**.
- Speaker 2 is a/an **{Speaker 2's Age Group}**.
- The social distance between the speakers is **{Social Distance}**.
- The power distance between the speakers is **{Power Distance}**

Question: Can this conversation happen in the real world? Answer with "Yes" or "No".

Figure 2: The prompt template for situational frame classification.

school life at a restaurant"). In general, we propose to train a frame classification model, along with several hand-written rules to filter out invalid frames. The process of this can be broken down into three steps: *Training Data Construction, Model Training, and Frame Classification*.

Training Data Creation. The training data of the frame classification model will have two parts, golden-labeled data and pseudo-labeled data. For the golden-labeled subset, we utilize the human-labeled frames from SocialDial, as many of the factor values of our data are sourced from this dataset. The number of human-labeled frames is 6,433. Regarding the pseudo-labeled data, we first sample 100,000 combinations of factor values, then prompt ChatGPT[1] for labeling. The prompt template is illustrated in Figure 2. To minimize the label errors made by ChatGPT API, we derive the probabilities of generating the tokens "Yes" or "No" from the API. Specifically, frames with either of the two probability scores higher than 0.85 are kept and assigned with the corresponding labels, and the remaining frames are removed. In total, we created a frame classification dataset with 41,016 samples, in which 16,547 samples are labeled as valid.

Model Training. With the constructed training dataset, we opt for the RoBERTa architecture (Liu et al., 2019) for frame classification. Specifically, the *large* version of the pretrained model is used for fine-tuning. We randomly split the constructed dataset into a training and development subset, with a ratio of 8:2. Adam optimization (Kingma and Ba, 2014) is used for model training. The choices of values for hyperparameters, such as learning rate, batch size, and number of epochs, are tuned through grid search over the development subset.

Frame Classification. The fine-tuned RoBERTa model is applied for frame classification. To ensure the label quality, we kept only the frames that the model predicted with a 0.995 probability value of the positive class. Additionally, we also introduced 30 handwritten simple rules that are used to filter out invalid frames. These rules are represented as combinations of different values for social factors that are not considered relevant in the real world.

3.4 Norm Extraction

The norm extraction process is illustrated in Figure 3. Specifically, we include the filtered situational frames in the prompts to discover social norms with ChatGPT. The prompt template includes four distinct parts:

- A template header describing the nature of the situational frame data.
- The body of the prompt template that outlines the social factors in a situational frame.
- A direct question describing the task of social norm extraction. This is followed by several constraints to ensure the quality and format of the generated norm statements are unified and controllable.
- Some Rules of Thumbs (RoTs) constraints. These contain RoT templates (Forbes et al., 2020) that will help to better structure the norm statement (eg. "In [X] culture, it is good to do action [Y], under situation [Z].").

3.5 Norm Deduplication

As the extracted norms can overlap in a single situational frame as well as across different frames, we remove one norm statement from each duplicating pair. This process is done separately for each culture. Specifically, we calculate the cosine similarity scores for every pair using their BERT embeddings (Devlin et al., 2019). If the similarity score is higher than 0.95, we flag the norm pair as duplicated.

3.6 Norm Verification

With the distinct norms obtained after the deduplication process, we begin to filter invalid norms. Invalid norm statements are norms that are incorrect in a specific culture, and we utilize ChatGPT for this verification process. Similar to Section

[1]https://openai.com/blog/chatgpt

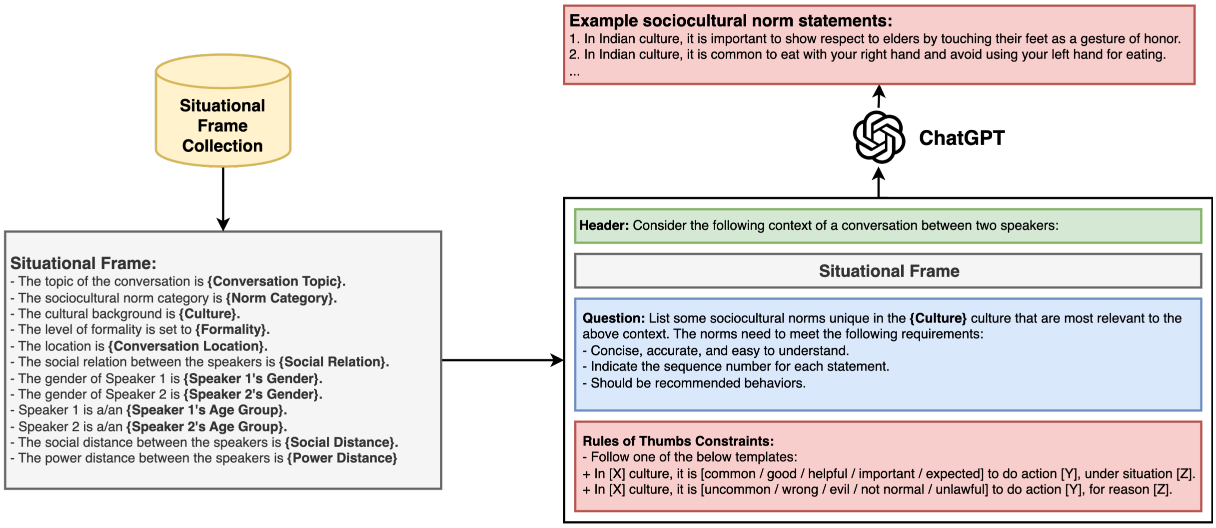

Figure 3: The norm extraction process with ChatGPT.

Culture	# of Norm Statements	# of Frames
American	27,481	4,505
Canadian	25,726	5,072
British	34,213	5,133
Chinese	24,789	4,496
Indian	25,760	4,675
Afghan	17,960	4,923
All	**155,929**	**28,804**

Table 1: Statistics of norms in different cultures.

3.3, we prompt ChatGPT with a Yes-No question, and derive the probability of the token "Yes" for filtering. Details of the prompt are given in the Appendix A.1. The probability threshold for valid norms is set to be 0.85.

3.7 Dataset Summary

With the above pipeline, we obtained the Multicultural Norm Base (MNB), which consists of 155,929 norm statements, extracted from more than 28,804 situational frames of 6 distinct cultures. The norm statements also represent real-world scenarios, where they reflect daily conversational situations through various speaker attributes. The norm statistics of the 6 cultures are reported in Table 1. The cultures have roughly equal numbers of situational frames. On average, about 5 norm statements are extracted with each situational frame in our data.

4 Experiments

To demonstrate the quality of our proposed method and dataset, we carry out experiments with our data and other related datasets. Our experiments are divided into two types: **Intrinsic Evaluation** and **Extrinsic Evaluation**. For intrinsic evaluation, we examine the quality of the constructed norm knowledge base and the norm extraction method. In the case of extrinsic evaluation, we demonstrate the applicability of our proposed dataset across different downstream tasks and compare the performance with other datasets.

4.1 Intrinsic Norm Discovery Evaluation

Similar to NormSage (Fung et al., 2023), we assess each norm statement on a Likert scale ranging from 1 to 5, where 1 denotes "Very Unsatisfied" and 5 denotes "Very Satisfied", for five criteria: *Relevance*, *Well-Formedness*, *Correctness*, *Insightfulness*, *Relatableness*. A detailed description of each criterion is provided in Appendix A.2.

As there are many norm statements in the dataset and evaluating all of them will be very time-consuming, we sample 100 norms from each culture for evaluation. Specifically, we randomly sample 100 situational frames from each culture and then sample 1 norm statement from each of the frames. This ensures that the selected data is diverse and covers a wide range of scenarios. To perform the evaluation, we employed 6 native human annotators, one in each of the 6 cultures to assess the data (e.g. a British person will label the British samples) to ensure the annotation quality.

Figure 4 illustrates the score distributions of different evaluation criteria. Regarding the distributions of *Well-Formedness*, the results are significantly high, as the percentage of the maximum

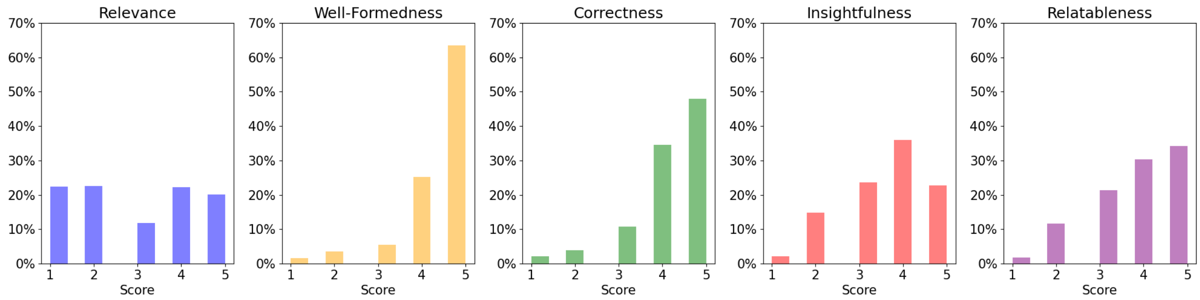

Figure 4: Score distributions of human annotators for different metrics.

score of 5 is above 60%. This shows that including the RoTs constraints in the prompt leads to a better structure of norm statements. For *Correctness* and *Insightfulness*, most score values also fall in the satisfaction categories (4 and 5), which highlights the quality of our norms. However, we observed a weakness of our method on the criteria *Relevance* and *Relatableness*, especially the distribution of *Relevance*, which seems to be evenly spread across all values. This indicates that ChatGPT is not utilizing all of the provided social factors for norm generation, which leads to lower *Relevance* scores and higher *Relatableness* scores. In other words, although the norms are correct and insightful, many of them are general and can be applied in various situations, which limits the diversity of the norm database. Upon closer inspection, we noticed that in many cases, ChatGPT only uses some of the factors in the prompt. We will describe some norm statements along with their problems in Appendix A.3 to further illustrate this observation.

4.2 Extrinsic Evaluation on Downstream Tasks

To set up extrinsic evaluations, we derive several related datasets and their corresponding downstream tasks, which can be categorized into generation tasks and classification tasks. For all extrinsic experiments, we will use Llama 2 (Touvron et al., 2023) and perform fine-tuning with different instruction tasks. Specifically, the 7B version of the Llama2-Chat model (Llama2-Chat-7B) is used for fine-tuning, as it already has been fine-tuned with a large set of instruction tasks and can be used as the baseline in experiments.

4.2.1 Generation Task

In terms of the generation task, we opt for the Moral Integrity Corpus (MIC) (Ziems et al., 2022) for our experiments. The norms covered in this dataset mostly are sourced from Reddit and belong to the American culture. The authors of MIC have set up the task of RoT generation, which requires models to generate a norm statement with a given dialogue content. To carry out the experiments, we compare the performance of the following models:

Metric	Llama2	Llama2 $_{SC}$	Llama2 $_{MNB}$
ROUGE-1	15.53	20.15	**21.56**
ROUGE-2	3.59	6.01	**7.92**
ROUGE-L	14.65	19.46	**20.48**
BLEU	11.95	16.16	**16.61**
BERT-Score	88.60	89.35	**90.94**

Table 2: Experimental results on the MIC dataset.

- **Llama2** The original Llama-2-Chat-7B model.
- **Llama2 $_{SC}$** The Llama-2-Chat-7B model fine-tuned with the SOCIAL-CHEM-101 dataset. The instruction task is generating a norm statement based on a given situation and a behavior.
- **Llama2 $_{MNB}$** The Llama-2-Chat-7B model fine-tuned with our MulticulturalNormBase dataset. The instruction task is to generate a norm statement based on a set of social factors (similar to how we extract the norms with ChatGPT in §4.4).

While the NormBank dataset can be used for training as it is also a norm dataset, its norms have a very different structure compared to our data as well as SOCIAL-CHEM-101 and MIC. The situational norms in NormBank are represented as taxonomies of various factors, while in the other 3 datasets, the norms are stated as Rules of Thumb statements. As converting the taxonomy-based norms into RoT involves great complexities, we chose to not experiment with the NormBank dataset for this generation task.

Following the authors of MIC, for the evaluation metrics, we apply the standard ROUGE (Lin and Hovy, 2003) (ROUGE-1, ROUGE-2, and ROUGE-L), BLEU score (Papineni et al., 2002), and BERT-Score (Zhang et al., 2020). The experimental results are reported in Table 2. All three models are evaluated in a zero-shot setting, meaning that they have not seen or been trained with the MIC dataset. It can be observed that when trained with cultural or commonsense knowledge data, the performance improves significantly over the baseline. Both the Llama models trained with SOCIAL-CHEM-101 and our dataset present better results than those of the baseline model. On all metrics, the model trained with our data (**Llama2**$_{\textbf{MNB}}$) achieves slightly higher results than the one trained with SOCIAL-CHEM-101 (**Llama2**$_{\textbf{SC}}$). This demonstrates that the extracted cultural norms are highly useful, and can be used to train models to adapt on different benchmarks.

4.2.2 Classification Tasks

Regarding the classification tasks, we consider the following datasets for evaluation:

EtiCor. (Ziems et al., 2023) This is a corpus of etiquettes, consisting of texts about social norms from five different regions across the globe, serving as a benchmark for evaluating LLMs for knowledge and understanding of region-specific etiquette. Specifically, the dataset covers 5 regions: *EA (East Asia)*, *IN (India)*, *MEA (Middle East & Africa)*, *NE (North America & Europe)*, and *LA (Latin America)*. With this data, the corresponding evaluation task is "Etiquette Sensitivity". Given a statement about etiquette, the task is to predict whether the statement is appropriate for a region. For this dataset, we use the entire data for evaluation.

NormBank. (Ziems et al., 2023) This is a knowledge base of situational norms. While the work of NormBank does not primarily concentrate on multicultural aspects, the dataset itself includes multicultural information. To extract the cultural information of norms in this dataset, we identify constraints that mention "Person Y's country is XX" and link them to specific cultures. We follow their evaluation on the task of "Norm Classification". Specifically, this task requires models to classify a combination of behavior and some constraints to be either *expected*, *okay*, or *unexpected*. To perform an evaluation on this dataset, we randomly split the samples into a training and test subset, with a ratio of 8:2. The reason for this split is that the models will be evaluated on the test subset, where the training set will be used to train a Llama 2 model, which will be used to compare with one trained with our proposed dataset.

Regarding the models for evaluation, we fine-tuned the Llama 2 model separately with the NormBank dataset and our dataset. Both models are trained with the classification task and the training procedure is different for each of the datasets, as their data attributes are different:

- **Llama2**$_{\textbf{NB-CLS}}$ The Llama-2-Chat-7B model fine-tuned with the training subset that we derived from the NormBank dataset. The model is trained for the task of norm classification, which utilizes the 3-class labels described previously.
- **Llama2**$_{\textbf{MNB-CLS}}$ The Llama-2-Chat-7B model fine-tuned with our MulticulturalNormBase dataset. The instruction task is also norm classification. Since the norms of our dataset are all recommended behaviors, we perform data augmentation to negate a portion of the data. Specifically, we apply two augmentation methods: rule-based and negative claim generation. For the rule-based method, we simply negate the adjectives that are defined in the norm statement template (eg. "It is bad/wrong/evil to do something" is converted to "It is good to do something"). In terms of the negative claim generation method, we utilize a pretrained BART model[2] to generate the negative version of a norm statement.

Apart from the fine-tuned models, we also experimented with a RAG (Retrieval Augmented Generation) based method with our data and the NormBank dataset. We derive two models - **Llama2**$_{\textbf{MNB-RAG}}$ and **Llama2**$_{\textbf{NB-RAG}}$ - which use the baseline Llama 2 model and retrieve the most relevant norms from our data and NormBank for a test sample, respectively. To ensure this method gets maximized results, we experimented with several numbers of norms being retrieved, ranging from 1 to 10, and reported only the best results. Interestingly, both **Llama2**$_{\textbf{MNB-RAG}}$ and **Llama2**$_{\textbf{NB-RAG}}$ achieve optimal results when using only 1 norm in the context.

Results on EtiCor. The experimental results on the EtiCor dataset are described in Table 3. The model trained with our dataset (**Llama2**$_{\textbf{MNB-CLS}}$) consistently demonstrates better results than the other two models, in all regions. The model shows the smallest absolute and relative improvements

[2] https://huggingface.co/minwhoo/bart-base-negative-claim-generation

Region	Llama2 (Baseline)		Llama2$_{\text{NB-CLS}}$		Llama2$_{\text{NB-RAG}}$		Llama2$_{\text{MNB-CLS}}$		Llama2$_{\text{MNB-RAG}}$	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
EA	68.08	69.97	63.40	66.88	63.61	63.67	**67.57**	**76.99**	62.16	73.75
IN	70.55	70.98	68.51	69.62	68.02	67.56	**74.65**	**80.72**	63.38	73.30
MEA	69.32	71.03	64.83	69.11	66.85	67.82	**69.74**	**78.94**	63.17	73.69
NE	76.43	82.62	77.94	84.07	73.21	79.40	**87.62**	**92.27**	76.28	84.95
LA	**66.89**	67.66	61.03	63.87	66.15	66.01	65.64	**76.05**	61.07	72.38
All	70.25	72.45	67.14	70.71	67.57	68.89	**73.04**	**80.99**	64.78	75.31

Table 3: Results of different models on the EtiCor dataset.

Culture	Llama2 (Baseline)		Llama2$_{\text{NB-CLS}}$		Llama2$_{\text{NB-RAG}}$		Llama2$_{\text{MNB-CLS}}$		Llama2$_{\text{MNB-RAG}}$	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
British	19.00	7.22	**43.21**	**38.26**	23.31	20.44	27.66	23.16	22.65	19.24
Canadian	9.06	5.17	**61.07**	**57.82**	26.49	32.23	33.62	35.51	13.81	16.07
American	14.01	4.67	**52.71**	**50.20**	16.68	15.69	32.76	32.60	19.77	19.89
Afghan	5.98	4.37	**69.18**	**67.27**	6.31	5.69	42.94	48.90	11.36	15.21
Indian	42.36	26.21	**51.35**	**45.28**	42.73	35.76	45.17	36.82	29.69	26.60
Chinese	31.86	16.23	**50.01**	**43.81**	33.33	25.24	37.50	27.93	32.52	26.60
All	18.60	9.68	**56.81**	**52.67**	23.73	21.09	37.30	35.71	22.38	21.20

Table 4: Results of different models on the NormBank dataset.

on the EA (East Asia) subset of EtiCor. This is because while our dataset consists of norms for the Chinese culture, EtiCor itself does not include Chinese data in the EA subset. Regarding **Llama2$_{\text{NB-CLS}}$**, while the nature of NormBank is also similar to EtiCor, however, the model does not achieve better overall results than the baseline Llama2 model, except for the NE (North America & Europe) subset, where the model demonstrates a significant improvement. This is understandable, as the portion of North American data accounts for almost 30% of the NormBank dataset. Despite being not as good as fine-tuning, the retrieval-based method also shows its improvements over the baseline, where the **Llama2$_{\text{MNB-RAG}}$** model achieves roughly 2.8% F1 improvement over the **Llama2** model.

Results on NormBank. The experimental results of different models on the NormBank dataset are described in Table 4. **Llama2$_{\text{NB-CLS}}$** obviously achieves the best results in terms of F1 and accuracy, as it is trained on the NormBank data. However, **Llama2$_{\text{MNB-CLS}}$** - the model trained with MNB still shows great improvements over the baseline, with more than 18% and 26% absolute improvements in overall accuracy and F1, respectively. Notably, **Llama2$_{\text{MNB-CLS}}$** demonstrates significant improvements over the baseline model on the Afghan culture - a culture that is considered low-resource. Furthermore, the **Llama2$_{\text{MNB-RAG}}$** model also outperforms the baseline and the **Llama2$_{\text{NB-RAG}}$** model on this Afghan subset. In terms of retrieval-based model, **Llama2$_{\text{MNB-RAG}}$** and **Llama2$_{\text{NB-RAG}}$** achieve competitive results, even though **Llama2$_{\text{NB-RAG}}$** takes advantage of retrieving norms from NormBank itself. Interestingly, **Llama2$_{\text{MNB-RAG}}$** reaches a better accuracy and F1 score than **Llama2$_{\text{NB-RAG}}$** on the American subset, despite this is the largest subset of the NormBank dataset. These results have proven that models utilizing our MNB dataset can generalize well across different domains and cultures, in both cases of fine-tuning and RAG.

5 Conclusions

In this paper, we propose an automatic norm discovery pipeline using ChatGPT for the multi-cultural setting. The pipeline extracts norm statements upon situational frames filled with crucial social factors. As real dialogues are not always available and can be limited to some domains, we have showcased that it is possible to extract meaningful norm statements only from social factors. Our derived norm database has shown its effectiveness in the experiments, achieving remarkable results on several downstream tasks and outperforming other norm datasets. In the future, we plan to expand the data with coverage to more cultures and implement large language models embedded with explicit cultural knowledge.

Limitations

Our proposed pipeline is based on the implicit knowledge of ChatGPT from OpenAI to extract cultural norm statements from conversational situations. While ChatGPT is trained on a large amount of data, its cultural knowledge and reasoning capabilities can have potential bias. We also acknowledge that cultural norms can vary and evolve significantly over time, which requires LLM to have better adaptation to new data. Despite the availability of more robust LLMs, such as GPT-4[3], we opted to use ChatGPT in our experiments due to the time limitation and costly usage of GPT-4. Additionally, more datasets should be compared with the proposed MNB dataset in future works. NormSage (Fung et al., 2023) is the closest work to ours, as it also has the multi-cultural element, but at the time of submitting this paper, the NormSage dataset and code are not publicly available for us to make a fair comparison in the experiments.

Another limitation of our work is the limited number of human annotators for intrinsic evaluation. We acknowledge that hiring more people to annotate the norms will better represent the norm quality, but due to the time constraint and cost limit, there is only one annotator for each culture. Although the chosen annotators are all native, there can still exist potential biases in the evaluation process.

Ethical Considerations

We recognize that automatically generated socio-cultural norm statements can carry an authoritative and normative tone (Fung et al., 2023). Therefore, we want to emphasize that these statements are not intended to serve as the basis for establishing a normative system or framework within any society. Their application in any operational system must be approached with caution. It is imperative to involve manual oversight to validate their accuracy prior to deployment. Consequently, these norm statements primarily serve only research purposes.

References

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

[3]https://openai.com/gpt-4

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. 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 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.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:6384–6392.

Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. Cskg: The commonsense knowledge graph. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 680–696. Springer.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Xuansong Li, Stephanie Strassel, Karen Jones, Brian Antonishek, and Jonathan G. Fiscus. 2022. Havic med novel 1 test – videos, metadata and annotation. *Web Download*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. Glucose: Generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3027–3035.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In *SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2712–2722, New York, NY, USA. Association for Computing Machinery.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. 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.

A Appendix

A.1 Norm Verification

As discussed in Section 3.6, we prompt ChatGPT to filter invalid norm statements. Figure 5 illustrates the prompt template for norm verification. Similar to previous prompt templates in Section 3.4 and Section 3.6, this template includes a header describing the nature of the situational frame, and a body outlining the social factors.

Header: Consider the following context of a conversation between two speakers:

Situational Frame:
- The topic of the conversation is **{Conversation Topic}.**
- The sociocultural norm category is **{Norm Category}.**
- The cultural background is **{Culture}.**
- The level of formality is set to **{Formality}.**
- The location is **{Conversation Location}.**
- The social relation between the speakers is **{Social Relation}.**
- The gender of Speaker 1 is **{Speaker 1's Gender}.**
- The gender of Speaker 2 is **{Speaker 2's Gender}.**
- Speaker 1 is a/an **{Speaker 1's Age Group}.**
- Speaker 2 is a/an **{Speaker 2's Age Group}.**
- The social distance between the speakers is **{Social Distance}.**
- The power distance between the speakers is **{Power Distance}**

Norm Statement:
- The norm statement corresponding to the given situation is as follows: **{Norm Statement}.**

Question: Is this a correct/acceptable socio-cultural norm in the given situation? Answer with “Yes” or “No”.

Figure 5: Prompt template for norm verification.

A.2 Intrinsic Evaluation

The definition for each criterion of the intrinsic evaluation process is as follows:

- **Relevance**. This criterion measures how well the situation inspires the generated norm. If a norm does not use the provided information from the situational frame, regardless of whether the norm is correct or not, the relevance score should be low.

- **Well-Formedness**. This criterion measures how well is the norm structured – is the norm self-contained, and does it include both a judgment of acceptability and an action or societal/cultural phenomena that is assessed?

- **Correctness**. This criterion measures the correctness of the norm. If a norm is considered to be correct in a given culture, its correctness score should be high.

- **Insightfullness**. This criterion measures the degree to which the norm conveys an enlightening understanding of what is considered acceptable and standard in the provided cultural background.

- **Relatableness**. This criterion measures the degree of generalization of a norm. If the given norm is highly applicable in various situations, the relatableness score should be high.

A.3 Problematic Norm Examples

Some examples of problematic norms are described in Table 5. We split the problems into three types: (i) generated norms that utilize only some of the speaker-related factors; (ii) generated norms that utilize only some of the conversation-related factors; and (iii) generated norms that do not use any of the given social factors. The examples shown here are the norms that are considered correct, but not relevant to their given situation frames. Hence, their *Relevance* scores are low and *Relatableness* scores are high, indicating that these norm statements are very general.

Category	Topic	Social Relation	Social Distance	Power Distance	Norm	Relevance	Relatableness
Generated norms that utilize only some of the speaker-related factors (social relation, social distance, and/or power distance)							
Apology	Sale	Elder-Junior	Working Relationship	Equal	In American culture, it is important to give equal opportunities for each speaker to express their opinions during a conversation.	1	5
Persuasion	Life-trivial	Peer-Peer	Friend	High	In American culture, it is expected to maintain a friendly and casual tone during peer-to-peer interactions.	2	4
Criticism	Food	Peer-Peer	Working Relationship	Equal	In Chinese culture, it is common to exchange business cards when meeting new colleagues.	1	4
Generated norms that utilize only some of the conversation-related factors (norm category and/or conversation topic)							
Request	Sale	Chief-Subordinate	Working Relationship	High	In American culture, it is common for customers to be given the option to return or exchange items within a specified time frame.	2	4
Greetings	Life-trivial	Peer-Peer	Friends	Equal	In Afghan culture, it is common to offer tea or refreshments when hosting guests.	1	5
Request	Food	Customer-Server	Stranger	Low	In Chinese culture, it is helpful to share dishes among the table instead of ordering individual portions.	3	4
Generated norms that do not use any of the given social factors							
Request	Life-trivial	Elder-Junior	Family	High	In Chinese culture, it is important to maintain a casual and friendly tone during an informal conversation.	2	4
Greetings	School Life	Peer-Peer	Stranger	Equal	In Indian culture, it is important to maintain a modest and appropriate dress code during professional interactions.	2	4
Criticism	Office Affairs	Chief-Subordinate	Working Relationship	Low	In British culture, it is important to be punctual and respect others' time commitments.	1	4

Table 5: Some norm examples with low *Relevance* score and high *Relatableness* score