Beyond English: Evaluating Automated Measurement of Moral Foundations in Non-English Discourse with a Chinese Case Study

Calvin Yixiang Cheng[1], **Scott A. Hale**[1,2]

[1]Oxford Internet Institute, University of Oxford, 1 St Giles', OX1 3JS, Oxford, UK
[2] Meedan, 575 Market St., 94105, San Francisco, USA
calvin.cheng@oii.ox.ac.uk, scott.hale@oii.ox.ac.uk

Abstract

This study explores computational approaches for measuring moral foundations (MFs) in non-English corpora. Since most resources are developed primarily for English, cross-linguistic applications of moral foundation theory remain limited. Using Chinese as a case study, this paper evaluates the effectiveness of applying English resources to machine translated text, local language lexicons, multilingual encoder-only language models, and decoder-only large language models (LLMs) in measuring MFs in non-English texts. The results indicate that machine translation and local lexicon approaches are insufficient for complex moral assessments, frequently resulting in a substantial loss of cultural information. In contrast, language models demonstrate reliable cross-language performance with transfer learning, with LLMs excelling in terms of data efficiency. Importantly, this study also underscores the need for human-in-the-loop validation of automated MF assessment, as even the most advanced models may overlook cultural nuances and face potential risks in cultural misalignment. The findings highlight the potential of LLMs for cross-language MF measurements and other complex multilingual deductive coding tasks.

Code —
https://github.com/calvinchengyx/cross-lan-mft-measure

Introduction

Moral intuitions have long fascinated social scientists, as they help explain a wide range of cognitive and behavioral phenomena across individuals and groups (Effron and Helgason 2022). Moral foundation theory (MFT) is among the most prominent psychology frameworks for understanding the origin and development of human morality (Graham et al. 2013). Rooted in moral nativism, MFT argues there are five universal moral foundations—care/harm, fairness/cheating, authority/subversion, loyalty/betrayal, and sanctity/degradation—that transcend languages and cultures and underlie people's moral judgments and decision-making processes (Graham et al. 2013).[1] While some scholars propose other foundations (Haidt 2012; Atari et al. 2023a), these five foundations have received the most empirical validation across domains, languages, and cultures (Iurino and Saucier 2020).

A growing body of literature employs MFT to investigate online social behaviors. For example, MFT provides a framework for understanding rising political polarization. Individuals who prioritize loyalty, authority, and sanctity foundations are more likely to endorse conservative views and engage in polarized political discourse, while those affiliated with liberal ideologies tend to value all MFs more evenly (Koleva et al. 2012; Haidt and Graham 2007). MFT also sheds light on online cultural clashes. Atran (2007) identified differing valuations of the sanctity foundation as a key factor in many religious and ideological conflicts. Beyond polarization, MFT has been applied to a range of social issues, including climate change (Markowitz and Shariff 2012), vaccine hesitancy (Amin et al. 2017), anti-abortion views (Koleva et al. 2012), nationalism (Kertzer et al. 2014), collective violence (Nussio 2023), and terrorism (Tamborini et al. 2020).

Given its broad relevance, measuring MF values in online discourse is essential; yet, automated extraction of MF values from large-scale texts remains challenging, particularly for non-English corpora. Like other latent human values, MFs are often conveyed through abstract narratives that vary across languages. Also, most computational resources for the measurement of MFs are designed for English content (Hoover et al. 2020; Trager et al. 2022). This reliance on English hinders cross-cultural comparative research on MFT and limits theoretical advancement from non-English data (Cheng and Zhang 2023). Although MFT is intended to apply across languages and cultures, the limited availability of non-English resources severely restricts its research scope and further development of the theoretical framework (Graham et al. 2013).

In this work, we investigate various computational approaches for cross-language measurement of MFs with a particular focus on data-efficiency. We use Chinese as an example and find that (1) MF local language lexicons yield suboptimal performance. They are worse than machine translation approaches that utilize established English measurements such as Mformer. (2) Multilingual encoder-only models can achieve moderate success when trained on annotated English data along with some local language la-

[1]Each MF includes both virtue and vice dimensions. We use the virtue label to represent the foundation value.

beled data. However, this strategy is less data-efficient in measuring MFs than in other deductive coding tasks, such as hate speech detection (Röttger et al. 2022). (3) Decoder-only LLMs outperform other approaches in cross-language MF measurements in both accuracy and data efficiency. Simply fine-tuning and augmenting with English annotated data can achieve strong performance on non-English corpora; (4) Nevertheless, this performance is inconsistent on different MF values, as LLMs may overlook cultural nuances in cross-language measurements, particularly for culturally distinct values

Related Work

The unique link between word usage and the expressed moral values provides a theoretical foundation for automated MF measurement from online texts (Brady, Crockett, and Van Bavel 2020; Gantman and Van Bavel 2014, 2016). Scholars have explored different computational approaches, including dictionaries (Graham, Haidt, and Nosek 2009; Hopp et al. 2021), word embeddings (Kwak et al. 2021; Araque, Gatti, and Kalimeri 2020), machine learning (Lan and Paraboni 2022), deep learning language models (Preniqi et al. 2024; Nguyen et al. 2024) and LLMs (Rathje et al. 2024). These computational methods demonstrate great advantages on scalability and labor intensity compared to traditional human annotations.

However, these approaches are primarily developed for English, and are not directly applicable to non-English texts due to several major concerns: differences in cultural contexts, the lack of annotated datasets, and limitations in domain generalizability. To address the cross-language challenges and bridge knowledge gaps in MF measurements, various computational approaches have been proposed, which can be broadly categorized into two paths: machine translation to English and the development of cross-language measurement tools (Zhuang et al. 2020).

English Centric Machine Translation

Translation is a widely used technique in cross-language MF measurement. For instance, the MFT survey has been translated into over 20 languages for cross-lingual studies (Yilmaz et al. 2016; Nilsson and Erlandsson 2015). For large scale text analysis, the development of multilingual neural machine translation has significantly improved translation quality compared to earlier statistical methods, enhancing context understanding, ambiguity resolution, and fluency (Stasimioti et al. 2020). This advancement enables a machine-translated approach to cross-language MF measurement by translating target languages into English and applying established English-based methods (Artetxe, Labaka, and Agirre 2020).

Established methods for automatically measuring MFs in English include dictionaries (Graham, Haidt, and Nosek 2009; Frimer et al. 2019; Hopp et al. 2021), word embeddings (Kwak et al. 2021; Araque, Gatti, and Kalimeri 2020), machine learning and deep learning models (Preniqi et al. 2024; Nguyen et al. 2024) trained on annotated English-language social media data (Hoover et al. 2020; Trager et al. 2022).

Moral foundation dictionaries (MFDs) Word-count methods with crafted English moral lexicons are common. There are four common English MFDs. The original MFD is an expert-crafted dictionary containing a list of 600 words across five foundation values (Graham, Haidt, and Nosek 2009). Frimer et al. (2019) then expanded this vocabulary to MFD2 to over 2,000 words by automatically identifying similar words with word2vec word embeddings (Mikolov et al. 2013). Similarly, Araque, Gatti, and Kalimeri (2020) extended the original MFD to a MoralStrength dictionary with approximately 1,000 English lemmas based on WordNet synsets (Princeton 2010). Compared to MFD2, MoralStrength added a round of crowd-sourced ratings on the expanded lemmas. Hopp et al. (2021), however, curated a fully crowd-sourced dictionary named eMFD. It is different from previous expert-curated dictionaries for its layperson focus, contextual annotations, probability labeling and large vocabulary with 3,200 words.

Moral word embeddings Semantic similarity methods using embeddings are another approach. To address the limitations of word-count methods, such as context insensitivity and vocabulary coverage (Nguyen et al. 2024), scholars have introduced semantic similarity methods. For example, Kwak et al. (2021) proposed an embedding framework FrameAxis. It predefines a vector space of micro moral frames with two sets of opposing seed words. Target documents are then converted into vectors using word embedding models, and their MF values are determined by comparing to the micro-frames. This method has been used to extract MF values from various online texts (e.g., Mokhberian et al. 2020; Jing and Ahn 2021).

Moral language models Supervised classification models with annotated English-language training data have also been used. With advancements in language models and efforts to create human-labeled MF training datasets (Hoover et al. 2020; Trager et al. 2022), recent studies have demonstrated the potential of fine-tuning language models. For example, Preniqi et al. (2024) fine-tuned a BERT-based classifier MoralBert with large-scale annotated English data and achieved state-of-the-art performance. To address generalizability limitations in out-of-domain datasets (Liscio et al. 2022), Nguyen et al. (2024) proposed another language model Mformer, and reported superior performance compared to other established English methods in evaluations.

Although machine translation offers several advantages in cross-language MF measurement, including interpretability, scalability, efficiency and accessibility, it also faces significant limitations. First, translation quality varies across languages and domains (Ranathunga et al. 2023). In some low-resource languages like Tamil, machine translation often makes errors in translating domain terms, polysemous words, and contains repetitions for semantically similar terms (Ramesh et al. 2021). Second, it often fails to retain non-propositional information, such as emotional nuances. This can lead to a loss of emotions, toning down, or amplification across languages, introducing bias in subsequent analyses (Troiano, Klinger, and Padó 2020). Third, machine translation struggles to capture cultural elements, which is

a major concern for cross-cultural and comparative research (Haidt 2012). It often shows limited performance with rare or culture-specific words, idiomatic phrases, and metaphor recognition (Dorothy, Keiko, and Vladimir 2019). Thus, it remains unclear whether machine translation is a reliable method for measuring cross-language MFs.

Cross-language Measurement Tools

A second path is to develop cross-language tools, where scholars create computational MF resources tailored to local languages. Common cross-language measurements include local language dictionaries, task-specific encoder-only language models, and LLMs.

Local language dictionaries Due to the efficiency at scale and multilingual capabilities, local language dictionaries are widely used to estimate MF values from non-English texts (Hopp et al. 2021). Developing local language dictionaries generally involves three steps: (1) translating English dictionaries to target languages; (2) adding culturally specific and non-translatable vocabulary; and (3) validating with native speakers and local language corpora. Several extensive non-English MFDs have been developed and validated in Turkish (Alper et al. 2020), Japanese (Matsuo et al. 2019), Portuguese (Carvalho et al. 2020), and Chinese (Cheng and Zhang 2023). Despite the abovementioned advantages, the dictionary approach still faces the inherent limitations of general bag-of-words methods. In this paper, we use C-MFD2—a Chinese MFD—as a cross-language tool to evaluate a local language dictionary approach. We also test a semantic similarity approach using the FrameAxis architecture and cross-language word embedding models.

Multilingual encoder-only models To overcome the limitations of bag-of-words methods, literature has suggested machine learning and deep learning approaches. A primary challenge with these methods is the scarcity of annotated data in local languages for model training (e.g., Ji et al. 2024; Nguyen et al. 2024). Therefore, scholars have adopted transfer learning techniques that leverage English-annotated resources for cross-language classifier development. Two major transfer learning strategies are commonly proposed: (1) machine-translating annotated English-language data into local languages to train monolingual models (Schuster et al. 2019), or (2) using annotated English-language data to train multilingual encoder-only models (Barriere and Balahur 2020). Multilingual models have demonstrated strong performance in deductive coding tasks, such as sentiment analysis (e.g., Barriere and Balahur 2020) and hate speech detection (e.g., Röttger et al. 2022). Nevertheless, they also have shown language bias in morality classification tasks, as pretrained multilingual encoder-only language models often display distinct moral directions across languages (Hämmerl et al. 2023). This paper focuses on the second transfer learning strategy and evaluates multilingual encoder-only models for cross-language MF measurement.

Large language models (LLMs) The rise of decoder-only LLMs provides an alternative for cross-language MF measurement. LLMs show exceptional zero/few-shot learning capability, enabling them to directly label human values out-of-the-box, which is particularly valuable for tasks with limited human annotated data (Ziems et al. 2024). They also have demonstrated strong performance in measuring various human values, including moral reasoning tasks (Ziems et al. 2024; Agarwal et al. 2024). Notably, LLMs sometimes are not as good as specialized fine-tuned language models (Amin, Cambria, and Schuller 2023; Preniqi et al. 2024), which may be due to the lack of explicit, colloquial definitions of the target human values (Ziems et al. 2024). MFT's well-established conceptual framework may help address this limitation. Not only are LLMs pre-trained on rich MFT literature (e.g., Abdulhai et al. 2024), but MFT also offers clear guidance for crafting clear and effective prompts. Additionally, LLMs trained on vast multilingual data exhibit promising capabilities to handle cross-language measurements (Ahuja et al. 2023).

Despite the strengths in accessibility, efficiency, multilingualism, and reasoning, there are also some concerns in using LLMs' in cross-language MF measurement. First, LLMs exhibit a substantial degree of subordinate multilingualism, displaying proficiency in some languages but not others (Zhang et al. 2023), which has a strong correlation with the proportion of those languages in the pre-training corpus (Li et al. 2024). Second, there are potential language biases, particularly in human-value relevant coding tasks (Kirk et al. 2024; Yu et al. 2024). For example, non-English prompts are more likely to generate malicious responses compared to English prompts (Shen et al. 2024).

Third, different LLMs exhibit varying baseline moral tendencies (Ji et al. 2024). For instance, GPT-3's MF preferences align more closely with politically conservative individuals when minimal prompt engineering is used in zero-shot learning (Abdulhai et al. 2024). These moral tendencies, however, are sensitive to prompting, with different prompting strategies significantly influencing classification outcomes in MF measurements (Abdulhai et al. 2024; Agarwal et al. 2024).

Thus, it is unclear how LLMs performs in cross-language MF measurement tasks. This paper selects a cutting-edge, open-source model—Llama3.1 (Meta 2024) to evaluate on LLM approach.

Data

Table 1 shows details of the benchmarking and training datasets used in this work. We used three human annotated datasets—moral foundation vignettes (MFV), Chinese moral scenarios (CCS) and Chinese core values (CCV), to benchmark the performance of different cross-language MF measurement approaches. We also used three English annotated MFT datasets for model training and fine-tuning.

Moral Foundation Vignettes (MFV) MFV is a list of social behaviors constructed by psychologists based on MFT, describing the violations of specific moral values from a third-party perspective (Clifford et al. 2015). It has been widely validated and used to assess measures of moral judgment (e.g., Graham, Haidt, and Nosek 2009; Kivikangas et al. 2021; Ji et al. 2024). We used the MFV as an expert-

	MFV	CCS	CCV	EN	Total
care/harm	27	389	3,030	16,607	20,053
loyalty/betrayal	16	248	1,712	11,772	13,748
authority/subversion	25	331	1,278	13,176	14,810
fairness/cheating	12	259	1,225	16,292	17,788
sanctity/degradation	10	226	247	9,165	9,648
non-moral	0	0	0	28,988	28,988
Total	90	1,453	7,492	71,242	80,277

Table 1: Moral foundation annotated datasets used in this paper. MFV, CCS and 20% of CCV were used for benchmarking; EN and 80% of CCV were used for fine-tuning XLM-T and Llama3.1-8b language models.

crafted benchmark stimulus to evaluate the baseline performance of different moral foundation measurement approaches. Since the original MFV is in English, we followed a careful translation process to ensure cultural neutrality in Chinese. First, a native Chinese speaker translated the vignettes with minor modifications to preserve cultural appropriateness. Then, two additional native speakers reviewed the translations to assess whether the scenarios remained representative and meaningful in Chinese cultural contexts. This ensured that the translated vignettes could serve as a culturally neutral benchmark for cross-method comparability.[2]

Chinese Moral Scenarios (CCS) CCS is list of moral scenarios written by 202 native Chinese to describe their intuitive understandings of MF values (Cheng and Zhang 2023). This reverse-annotation method is commonly used for validating MF measurements (e.g., Cheng and Zhang 2023; Frimer et al. 2019; Matsuo et al. 2019). It incorporates culturally specific content, reflecting native speakers' natural and intuitive understanding of MFT in a real-word context.

Chinese Core Values (CCV) CCV is a human-annotated real-world dataset, including 6,994 sentences collected from four local news websites.[3] The dataset is annotated by three native Chinese speakers based on the Chinese core socialist moral value coding scheme, which is highly correlate with the five universal moral foundation values (Liu et al. 2022). The original CCV dataset includes eight labels,[4] we employed five Chinese native speakers with postgraduate degrees to re-label the action categories in CCV to five moral foundation values, excluding vice and virtue. The mapped values were determined by majority vote. We sample 20% of the CCV as the primary benchmarking dataset to evaluate the performance across approaches stratifying on the values. The remainder is used as training data to test the data-efficiency fine-tuning language models.

All benchmarking documents are single-class labeled. For multi-class predicted documents, we decided the measurement performance by a lenient evaluation criterion: a prediction is considered correct if one of the predicted values matches the true label.

English annotated data (EN) We use three English annotated MF datasets for transfer learning, including a Twitter corpus (Hoover et al. 2020), a Reddit corpus (Trager et al. 2022) and a news corpus (Hopp et al. 2021), which are widely used in training English MF classifiers and show reliable performance (Nguyen et al. 2024; Preniqi et al. 2024).

Methods

Machine Translation

We use Google Translate as an example of a machine translation approach due to its accessibility and consistent performance across domains. First, we machine translate benchmarking datasets except MFV to Chinese using the Google Cloud API—Basic Translation service. Then we estimate the MF values from translated documents with established English measurements, including lexicons MFD (Graham, Haidt, and Nosek 2009), MFD2 (Frimer et al. 2019), eMFD (Hopp et al. 2021) and MoralStrength (Araque, Gatti, and Kalimeri 2020); word embeddings with FrameAxis (Kwak et al. 2021); and specialized-fine-tuned language models MoralBert (Preniqi et al. 2024) and Mformer (Nguyen et al. 2024).

For MFD, MFD2, and eMFD, we calculate word frequencies using the eMFDscore Python package (Hopp et al. 2021).

For MFD and MFD 1.0, each document's MF value is determined by the most frequent MF class in the respective dictionary. A document is mapped to multiple classes if there is an equal number of class matches. If there are no matching words, no class is assigned to the document. eMFD, however, assigns probabilities to its vocabulary, representing their likelihood of being associated with certain MF classes. We sum the probabilities and label the document with the class that has the highest sum.

We use the moralstrength package (Araque, Gatti, and Kalimeri 2020) for the MoralStrength dictionary. We first test its performance of the lexicon features alone with bag-of-word methods; then we train a Support Vector Machine (SVM) model and combine its lexicon features. Since English training sets contain many non-moral labels, while the benchmarking dataset contains only moral labels, we train two SVM models: one with the full training data and another with only moral-labeled training data.

For word embedding methods, we use the FrameAxis Python package (Kwak et al. 2021) to compute anchor micro-frames based on different MFDs with the word2vec embedding model (Mikolov et al. 2013). For each MFD, we generate the corresponding micro moral frames from the vocabulary in its class. We then compute and aggregate word contributions to each microframe in the document, and label the document's moral class by identifying significant microframes through comparison with a null model.

[2]The translated vignettes are available on the project's GitHub page.

[3]CCV includes data from the CMOS corpus (Peng et al. 2021), China Cultural and Ethical Website `wenming.cn`, Youth Patriotism News `agzy.youth.cn`, and Sohu News `news.sohu.com`

[4]The Chinese core socialist moral values include civility, justice, equality, rule of law, patriotism, dedication, integrity, and friendship. The Appendix showed the curated mapping scheme.

For language models, we apply pre-trained language models MoralBert and Mformers from HuggingFace with their default settings and no additional fine-tuning.

Local Language Lexicons

We apply C-MFD2 (Cheng and Zhang 2023) to evaluate the performance of the local language dictionary approach. We test two techniques for locally-developed MF lexicons: word counts and embeddings. For the word embedding method, we test two approaches with the fastText model (Grave et al. 2018). One involves measure simple semantic similarity. Words in C-MFD2 are grouped into five pseudo-documents based on their MF labels, each serving as anchor frames. MF values are then determined by calculating the semantic distance between the text to be classified and these anchor frames.

The other uses FrameAxis to construct anchor frames. As C-MFD2 does not have the virtue/vice dimension, which is essential to calculate micro-frames in FrameAxis, we automatically assign this dimension using the RoBERTa-based Chinese sentiment model c2-roberta-base-finetuned-dianping-chinese from HuggingFace.

Multilingual Encoder-only Models

We select *XLM-T* as the base model to test transfer learning on the multilingual encoder-only model approach (Barbieri, Anke, and Camacho-Collados 2022). Fine-tuned on 198 million multilingual tweets on the XLM-RoBERTa architecture—originally trained on 2.5 TB of Common Crawl data (Conneau et al. 2019)—this model is particularly well-suited for analyzing online content. Also, previous research indicates that it has strong performance in cross-language deductive coding tasks such as hate speech detection, compared to other encoder-only models (Röttger et al. 2022).

We follow Nguyen et al. (2024)'s experience on fine-tuning Mformer with some tweaks. First, we replace the base architecture from RoBERTa-base (Liu et al. 2019) with twitter-xlm-roberta-base and tokenization is handled by the model's built-in tokenizer (Barbieri, Anke, and Camacho-Collados 2022), with token sequences truncated to a maximum of 512. Second, we set the learning rate, epochs, and batch size to $2e-5$, 3, and 16 respectively following Röttger et al. (2022). Third, we opt to fine-tune five binary classifiers rather than a single multi-label classifier because binary models generally outperform multi-label models in English MF measurements (Nguyen et al. 2024).[5] Fourth, we adopt a conservative under-sampling strategy in the English annotated training dataset to address class imbalance, establishing a baseline for future improvements.

After fine-tuning with English annotated data, we further fine-tune each base model with the 80% of CCV dataset in order to test the data-efficiency as in Röttger et al. (2022). We incrementally train models with batches of additional CCV data, with each batch containing 100 annotated records.

[5]We note that in hate speech detection, Röttger et al. (2022) found no significant performance difference between binary and multi-label models when using XLM-T. Given that multi-label models require less storage and training resources, they are a viable alternative for future applications.

Large Language Models

For the decoder-only LLMs approach, we select the Llama3.1-8b instruct model, an open-source LLM developed by Meta with a reasonable balance between model performance and computational cost. Compared to closed-source models like GPT-4, Llama3.1 offers greater control, flexibility, transparency, and reproducibility—all of which are important in human-value measurement tasks.

We first test LLMs with prompt-engineering and few-shot learning. Next, we apply the same fine-tuning process used for XLM-T to Llama3.1-8b. Notably, an additional round of data augmentation is performed afterwards, where all English annotations are machine-translated into Chinese using the Google Translate API. Fine-tuning is conducted with the unsloth package using 4-bit quantization on a single NVIDIA L40S GPU. Additionally, we test data efficiency with 20 batches of Chinese annotated items from the CCV dataset. A conservative under-sampling strategy is applied as well with each batch containing 50 records evenly distributed across the five classes.

Qualitative Analysis

We conducted qualitative analysis to gain in-depth understanding of the cultural loss in moral foundation measurements across different approaches. By randomly sampling 100 mislabeled records, we compared predicted labels to ground truth (i.e., human annotated labels), focusing on MFs that are known with cultural distinctions between English and Chinese (i.e., authority, loyalty, and sanctity).

Results

Machine Translation

As shown in Table 2, the performance of machine translation generally fall short in evaluation. With the MFV benchmarking dataset, the lexicon method MFD2 shows the best performance with a weighted F1 score of 0.60. In the reverse-annotated CCS dataset, a simple SVM model with lexicon features outperforms other measurements ($F1 = 0.74$), but the model coverage is relatively low at only 23%. In the real-word CCV dataset, the deep learning model Mformer exhibits the strongest performance ($F1 = 0.47$) and maintains comparable results across the other two benchmark datasets. Note that although some MF classes in Mformer displayed good performance (i.e., care/harm, $F1 = 0.72$), some fine-grained MF measurements are very poor. For example, Mformer's prediction on "loyalty" ($F1 = 0.21$) is worse than random guessing baseline ($F1 = 0.23$) in the cross-language evaluation setting.

Local Language Lexicons

Local language lexicon approaches demonstrate similarly moderate performance with word-count methods. Table 3

MFV	Auth	Care	Fair	Loya	Sanc	Acc	Cov	Fw	Fm
Baseline	0.28	0.30	0.13	0.18	0.11	0.23	1.00	0.23	0.20
MFD	**0.77**	0.29	0.00	0.55	0.00	0.56	0.28	0.55	0.32
MFD2	0.69	0.50	**0.75**	**0.59**	0.33	**0.59**	0.46	**0.60**	**0.57**
eMFD	0.39	0.58	0.31	0.22	**0.44**	0.42	**1.00**	0.41	0.39
MS	0.00	0.00	0.00	0.33	0.31	0.17	0.20	0.13	0.13
FA+MFD	0.39	0.06	0.00	0.29	0.31	0.27	**1.00**	0.21	0.21
FA+MFD2	0.19	0.18	0.06	0.43	0.17	0.22	**1.00**	0.21	0.21
FA+eMFD	0.07	0.47	0.19	0.40	0.17	0.31	**1.00**	0.28	0.26
svm+MS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
svm+MS*	0.07	0.32	0.21	0.09	0.00	0.19	**1.00**	0.16	0.14
MoralBert	0.27	0.55	0.50	0.10	0.30	0.38	**1.00**	0.36	0.34
MFormer	0.65	**0.65**	0.72	0.31	0.40	0.58	**1.00**	0.57	0.55
CCS									
Baseline	0.23	0.27	0.18	0.17	0.16	0.21	1.00	0.21	0.22
MFD	0.47	0.62	0.67	0.18	**0.84**	0.54	0.6	0.54	0.55
MFD2	0.55	**0.71**	0.71	0.61	0.71	0.66	0.75	0.66	0.66
eMFD	0.48	0.52	0.48	0.44	0.25	0.47	0.98	0.45	0.43
MS	0.16	0.11	0.16	0.18	0.21	0.16	0.17	0.16	0.16
FA+MFD	0.45	0.38	0.31	0.26	0.47	0.38	**1.00**	0.38	0.37
FA+MFD2	0.39	0.57	0.57	0.39	0.43	0.47	**1.00**	0.47	0.47
FA+eMFD	0.33	0.42	0.28	0.17	0.22	0.30	**1.00**	0.30	0.28
svm+MS	**0.68**	**0.71**	**0.77**	**0.85**	0.61	**0.74**	0.23	**0.74**	**0.72**
svm+MS*	0.21	0.39	0.28	0.15	0.14	0.26	**1.00**	0.25	0.23
MoralBert	0.41	0.63	0.58	0.64	0.49	0.57	**1.00**	0.55	0.55
MFormer	0.65	**0.71**	0.70	0.71	0.71	0.69	**1.00**	0.69	0.70
CCV									
Baseline	0.18	0.40	0.16	0.23	0.03	0.27	1.00	0.27	0.20
MFD	**0.33**	0.40	0.43	0.3	0.00	0.34	0.47	0.35	0.29
MFD2	0.23	0.62	0.43	0.27	0.06	0.43	0.72	0.43	0.32
eMFD	0.11	0.59	0.36	0.20	0.02	0.40	**1.00**	0.36	0.26
MS	0.17	0.21	0.23	0.23	0.09	0.20	0.36	0.21	0.19
FA+MFD	0.29	0.30	0.27	**0.31**	0.06	0.28	**1.00**	0.29	0.25
FA+MFD2	0.07	0.49	0.35	0.30	0.11	0.34	**1.00**	0.34	0.27
FA+eMFD	0.04	0.56	0.31	0.24	0.08	0.38	**1.00**	0.34	0.25
svm+MS	0.28	0.59	0.49	0.27	**0.18**	0.47	0.17	0.44	**0.36**
svm+MS*	0.12	0.35	0.25	0.12	0.03	0.23	**1.00**	0.23	0.17
MoralBert	0.15	0.62	0.47	0.24	0.10	0.45	**1.00**	0.41	0.32
MFormer	0.23	**0.72**	**0.52**	0.21	0.14	**0.50**	**1.00**	**0.47**	**0.36**

Table 2: Established English moral foundation measurements applied to machine translated text. Acc, Cov, Fw and Fm refers to accuracy, coverage, F1 weighted and F1 macro respectively. Baseline, FA and MS in the first column represent random guessing, FrameAxis and MoralStrength. The best performing methods for each dataset are in bold. MoralStrength models trained with no non-moral labeled data are marked with a star (*). The Coverage column measures the percentage of all moral labels in the prediction.

shows that across all benchmarking datasets, C-MFD2 consistently outperforms other methods although its performance is generally still poor. In the real-world CCV dataset, C-MFD2 achieves the best performance ($F1 = 0.43$), which is comparable to the machine translation approach with Mformer. We find multilingual word embedding model fastText with FrameAxis framework do not improve the cross-language measurement performance in this task.

MFV	Auth	Care	Fair	Loya	Sanc	Acc	Cov	Fw	Fm
Baseline	0.28	**0.30**	0.13	0.18	0.11	0.23	1.00	0.23	0.20
cMFD2	**0.48**	0.18	**0.67**	0.29	**0.22**	**0.39**	0.40	**0.42**	**0.37**
FT	0.44	0.00	0.00	0.12	0.18	0.30	**1.00**	0.16	0.15
FT+FA	0.08	0.00	0.00	**0.30**	0.00	0.19	**1.00**	0.08	0.08
CCS									
Baseline	0.23	0.27	0.18	0.17	0.16	0.21	1.00	0.21	0.22
cMFD2	**0.74**	**0.76**	**0.73**	**0.68**	**0.74**	**0.74**	0.76	**0.73**	**0.73**
FT	0.44	0.32	0.35	0.39	0.37	0.39	**0.97**	0.37	0.37
FT+FA	0.14	0.05	0.15	0.29	0.02	0.20	**0.97**	0.12	0.13
CCV									
Baseline	0.18	0.40	0.16	0.23	0.03	0.27	1.00	0.27	0.20
cMFD2	0.26	**0.63**	**0.45**	0.28	**0.09**	**0.45**	0.64	**0.43**	**0.34**
FT	**0.28**	0.17	0.09	0.01	0.03	0.20	**0.98**	0.13	0.12
FT+FA	0.02	0.03	0.04	**0.37**	0.00	0.23	**0.98**	0.10	0.09

Table 3: The performance of C-MFD2 in cross-language MF measurement. FT and FA in the first column represent FastText and FrameAxis. The best performing methods for each dataset are in bold. Other abbreviations are the same as those in Table 2.

Multilingual Encoder-only Models

The multilingual encoder-only model XLM-T, trained on the same English annotated data, shows a moderate but reduced performance compared to the monolingual English model Mformer. In Table 4, XLM-T shows moderate performance across all benchmarking datasets ($F1_{MFV} = 0.71, F1_{CCS} = 0.71, F1_{CCV} = 0.63$) outperforming both machine translation and local lexicon approaches. In addition, its performance is generally consistent across the five foundation values, showing a better reliability in the fine-grained MF measurements. This consistency is likely due to training five separate classifiers. In the real-world CCV dataset, the fine-tuned XLM-T model has moderate performance in four out of five foundation values with an average F1 score of 0.63. The "authority" model has lower performance with $F1 = 0.46$.

In the follow-up evaluation of batch training with local language data, we observe improved model performance as the volume of local language training data increases. As shown in Figure 1, the fine-tuned XLM-T models achieve higher F1 scores with more locally labeled data. For example, the F1 score of the "loyalty" model rose from 0.62 to 0.80 with 22 batches of local-language data. Similar increasing patterns are observed across all five models.

However, the amount of annotated non-English data required to reach reliable classification thresholds for MFs is much more than for hate speech detection (Röttger et al. 2022). On average, 16 batches are needed to reach the F1 score of 0.70 and 49 batches to reach 0.80. The "care" model, which is the best performing model of the five, requires six additional batches to reach an F1 score of 0.70 and 27 batches to reach 0.80. There are 10 batches of data available for the "sanctity" model, and it shows no significant

MFV	Auth	Care	Fair	Loya	Sanc	XLM-T Avg
0	0.70	0.81	0.76	0.88	0.85	0.80
1	0.27	0.41	0.15	0.54	0.15	0.30
Acc	0.58	0.71	0.62	0.81	0.74	0.69
Fm	0.49	0.61	0.45	0.71	0.50	0.55
Fw	0.58	0.69	0.68	0.82	0.77	0.71
CCS						
0	0.66	0.76	0.79	0.91	0.89	0.80
1	0.19	0.44	0.02	0.54	0.51	0.34
Acc	0.52	0.66	0.65	0.85	0.82	0.70
Fm	0.42	0.60	0.4	0.73	0.70	0.57
Fw	0.55	0.67	0.65	0.85	0.83	0.71
CCV						
0	0.51	0.63	0.74	0.78	0.75	0.68
1	0.24	0.58	0.20	0.41	0.10	0.31
Acc	0.40	0.60	0.61	0.68	0.61	0.58
Fm	0.37	0.60	0.47	0.60	0.43	0.49
Fw	0.46	0.61	0.65	0.70	0.73	0.63

Table 4: Multilingual encoder-only models for moral foundation measurement. *XLM-T Avg* refers to the average F1 score across five foundation models.

improvement even once all 10 batches are used for training.

It is worth noting that limited local language annotated data may also negatively impact the model's performance. For example, the "fairness" model's F1 score initially drops from 0.50 to 0.46 when fine-tuned with local-language data. It only begins to stabilize and improve after 11 batches. In summary, while moderate performance can be achieved with local-language annotations, considerably more data is required to attain robust classification performance with the multilingual encoder-only model approach in cross-language MF measurements.

Large Language Models

Table 5 shows the results of the LLMs approach. With only prompt engineering, Llama3.1-8b already performs better than the machine translation and local language lexicon approaches across all benchmarking datasets. After fine-tuning with annotated English-language data, it immediately achieves strong performance on the MFV ($F1 = 0.81$) and CCS ($F1 = 0.82$) datasets, and moderate performance on the CCV ($F1 = 0.65$) dataset. This performance exceeds that of XLM-T with the same English-language training data ($F1 = 0.63$).

The model is fine-tuned with a data augmentation strategy where the English labeled data are machine translated into Chinese and fed into the LLM again. This step significantly improves the model performance and results in the best performance for all three benchmarking dataset (MFV $F1 = 0.86$; CCS $F1 = 0.82$; CCV $F1 = 0.74$). These F1 scores exceed all other approaches tested in this paper.

Additionally, we notice that English prompts generally outperform non-English prompts on Llama3.1-8b, even when analyzing non-English documents. This difference is more pronounced in the base instruct model but becomes less significant with progressive fine-tuning using annotated data. After fine-tuning with English data, the difference in the F1 scores for English and Chinese prompts on the CCV dataset decreases from 0.11 to 0.08, further reducing to 0.03 with data augmentation and 0.02 when fine-tuned with both English and translated Chinese data.

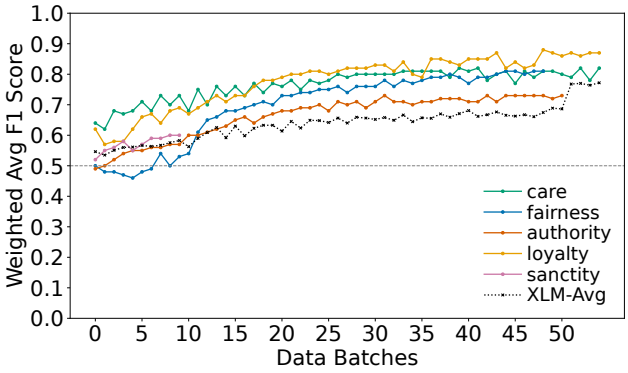

Figure 1: Accumulated fine-tuning of the XLM-T with local language annotated data from the CCV dataset. Each batch includes 100 items.

Moreover, a better LLM base model may further improve the model performance. Table 6 shows that under the same condition, Llama3.1-70b outperforms Llama3.1-8b when prompting in local languages. In the CCV dataset, under a few-shot learning setup with Chinese prompting, the larger LLM ($F1 = 0.60$) is better than both its smaller counterpart ($F1 = 0.42$) and the English prompting counterpart ($F1 = 0.51$).

Qualitative Analysis of Cultural Nuances

We selected MFormer and fine-tuned Llama3.1-8b-instruct with English and translated Chinese data as models to analyze cultural loss in translation and non-translation approaches. We identified three common types of information loss during the translation process (See Figure 4 in the Appendix). First, idioms and slang are often mistranslated, as in Example 1, where the Chinese phrase meaning "faked or staged accident for compensation" is simply translated to its superficial meaning. Second, contextual meaning. As illustrated in Example 2, where the English translation "look towards" misses the cultural connotation of "favoring someone," resulting in different expressed moral foundations. Third, political euphemisms lose their cultural context, as in Example 3, where the English translation fails to convey that the text refers to civil servant behavior. These information loss in the translation process often result in mislabeling in MF measurement.

We further analyzed the non-translational approach to understand how LLMs process cultural nuances. Compared to the translational approach, LLMs better recognize slang and idioms (Examples 2 and 13), but demonstrate more subtle cultural misalignment. Our analysis revealed several distinct patterns. Most notably, LLMs failed to distinguish certain

MFV	Auth	Care	Fair	Loya	Sanc	Acc	Cov	Fw	Fm
Baseline	0.28	0.30	0.13	0.18	0.11	0.23	1.00	0.23	0.20
en×∅	0.41	0.78	0.40	0.35	**0.40**	0.55	**0.61**	0.54	0.47
zh×∅	0.56	0.40	0.25	0.00	0.00	0.42	0.36	0.34	0.24
en×en	0.89	0.87	0.94	0.57	0.00	0.84	0.41	0.81	0.65
zh×en	0.82	0.88	0.75	0.67	0.00	0.79	0.32	0.77	0.62
en×zh	0.89	0.86	0.90	**0.80**	0.00	0.86	0.54	0.85	0.69
zh×zh	0.91	0.83	0.91	0.75	0.50	0.85	0.58	0.84	**0.78**
en×(en+zh)	0.88	**0.89**	**0.95**	**0.80**	0.00	**0.87**	0.60	**0.86**	0.70
zh×(en+zh)	**0.93**	0.88	0.91	0.77	0.00	**0.87**	0.60	0.85	0.70
CCS									
Baseline	0.23	0.27	0.18	0.17	0.16	0.21	1.00	0.21	0.22
en×∅	0.58	0.78	0.71	0.64	0.71	0.70	**0.93**	0.69	0.68
zh×∅	0.66	0.80	0.71	0.63	**0.72**	0.71	0.82	0.71	0.71
en×en	**0.78**	**0.85**	**0.88**	0.84	0.68	**0.82**	0.52	**0.82**	0.80
zh×en	0.75	0.84	**0.88**	0.80	0.69	0.80	0.45	0.80	0.79
en×zh	0.73	**0.85**	0.84	0.85	**0.72**	0.81	0.56	0.80	0.80
zh×zh	0.74	0.82	0.85	0.86	0.65	0.80	0.56	0.80	0.78
en×(en+zh)	0.76	**0.85**	0.86	**0.87**	0.69	**0.82**	0.55	**0.82**	**0.81**
zh×(en+zh)	0.74	0.84	0.85	**0.87**	0.70	0.81	0.57	0.81	0.80
CCV									
Baseline	0.18	0.40	0.16	0.23	0.03	0.27	1.00	0.27	0.20
en×∅	0.29	0.69	0.47	0.50	0.06	0.54	**0.80**	0.53	0.40
zh×∅	0.29	0.70	0.35	0.16	0.13	0.48	0.61	0.42	0.32
en×en	0.28	0.83	0.68	0.55	0.00	0.66	0.62	0.65	0.47
zh×en	0.19	0.82	0.59	0.23	0.00	0.59	0.41	0.57	0.37
en×zh	0.39	**0.85**	0.77	0.69	0.36	0.73	0.68	0.72	0.61
zh×zh	0.32	0.81	0.73	0.70	0.36	0.71	0.69	0.69	0.59
en×(en+zh)	**0.42**	**0.85**	**0.79**	**0.72**	**0.43**	**0.75**	0.72	**0.74**	**0.64**
zh×(en+zh)	0.36	0.84	0.75	**0.72**	0.31	0.73	0.72	0.72	0.60

Table 5: Llama3.1-8b model with few-shot learning prompts for cross-language moral foundation measurements. The first language refers to the language of prompt while the languages after × refer to the language(s) of fine-tuning datasets. ∅ denotes the empty set when no fine-tuning is done. For example, zh×(en+zh) denotes a Chinese prompt on a model fine-tuned using English and Chinese data. The best performing method per dataset is in bold. Column labels are consistent with Table 2.

Confucian morality concepts such as filial piety, which encompass family responsibilities to respect and care for elders (Examples 6, 8, 9, and 10). Native Chinese speakers tend to labeled these instances as "authority," whereas language models often only categorized them as "care." Additionally, we observed misalignment regarding government and local political entities. Human annotators tended to label cases involving government (Examples 3, 4) and officials (Example 11) as "loyalty," while LLMs classified them as "authority" or "fairness." Finally, examining misalignment in "sanctity" revealed that Chinese annotators frequently associated sanctity with corrupt politicians or role model civil servants (Examples 11 and 12), a distinction rarely captured by LLMs.

MFV	Auth	Care	Fair	Loya	Sanc	Acc	Cov	Fw	Fm
Baseline	0.28	0.30	0.13	0.18	0.11	0.23	1.00	0.23	0.20
8b en	0.41	0.78	0.40	0.35	0.40	0.55	0.61	0.54	0.47
8b zh	0.56	0.40	0.25	0.00	0.00	0.42	0.36	0.34	0.24
70b en	0.67	**0.82**	0.52	0.48	0.55	**0.66**	**0.83**	**0.66**	0.61
70b zh	**0.73**	0.67	**0.53**	**0.56**	**0.60**	0.63	0.64	0.64	**0.62**
CCS									
Baseline	0.23	0.27	0.18	0.17	0.16	0.21	1.00	0.21	0.22
8b en	0.58	0.78	**0.71**	0.64	0.71	0.70	0.93	0.69	0.68
8b zh	**0.66**	0.80	**0.71**	0.63	**0.72**	**0.71**	0.82	**0.71**	**0.71**
70b en	0.54	0.77	0.63	0.68	0.65	0.67	**0.96**	0.66	0.65
70b zh	0.56	**0.83**	0.63	**0.71**	0.70	0.70	0.89	0.69	0.69
CCV									
Baseline	0.18	0.40	0.16	0.23	0.03	0.27	1.00	0.27	0.20
8b en	**0.29**	0.69	0.47	0.50	0.06	0.54	0.80	0.53	0.40
8b zh	**0.29**	0.70	0.35	0.16	0.13	0.48	0.61	0.42	0.32
70b en	0.18	0.69	0.52	0.41	0.15	0.55	**0.86**	0.51	0.39
70b zh	0.18	**0.76**	**0.54**	**0.59**	**0.18**	**0.62**	0.63	**0.60**	**0.45**

Table 6: *Llama3.1* 8b and 70b models with few-shot prompting for cross-language moral foundation measurement. The best performing method for each dataset is in bold. Column labels are consistent with Table 2.

Discussion and Limitations

We find traditional approaches, such as machine translation and local language lexicons, may not be the most effective solutions for cross-language moral foundation measurements. Instead, fine-tuning multilingual encoder-only models and leveraging LLMs through transfer learning emerge as promising alternatives. Notably, LLMs demonstrate strong performance and data efficiency, making them particularly well-suited for addressing the challenges of cross-language MF analysis.

Above all, simple machine translation approach has proven suboptimal in cross-language MF measurement. This issue is particularly concerning for measuring culturally significant values. Machine translation often introduces bias in the rendering of slangs, contextual meanings, and political euphemisms into local languages. Moreover, when relying on pretrained English classifiers such as MFormer—which is predominantly trained on English-annotated data—this cultural information loss may be further amplified. For instance, MFormer's classification performance on the "loyalty" foundation in translated data ($F1 = 0.21$) falls below random chance ($F1 = 0.23$), indicating a substantial loss of cultural nuance. Nevertheless, it is important to note that machine translation models are rapidly evolving. Current deficiencies—such as loss of contextual nuance or culturally embedded moral cues—limit their effectiveness in this domain. Advances in context-aware and domain-adapted translation models (Jin et al. 2023; Saunders 2022), particularly those fine-tuned on moral discourse, may mitigate some of these issues, which deserves future exploration.

Researchers should also be cautious with local language lexicons—they perform worse than the machine translation

for cross-language MF measurement in some cases. Given the inherent limitations of lexicon-based methods and the extensive resources required to develop them, they are inefficient for cross-language MF measurement, particularly at the document level where semantic complexity is higher. However, culturally specific moral lexicons may still provide valuable insights at the vocabulary level for cross-cultural research, though broader applications should be approached with careful validations.

Fine-tuning LLMs is more data-efficient than fine-tuning multilingual encoder-only models for cross-language MF measurement. In the CCV benchmarking dataset, XLM-T requires on average more than 2,000 local-language annotated records (20 batches) to achieve strong performance ($F1 = 0.75$). In contrast, Llama3.1-8b can reach comparable performance with only English data machine-translated to Chinese and thus no local language labeling. Our findings also suggest that larger LLMs such as Llama3.1-70b may perform better for measuring MFs in non-English corpora. That is, updating the base LLM to a more power model with enhanced multilingual capabilities could further improve the performance of the LLM approach.

Nevertheless, several limitations of using LLMs for cross-language MF measurement should be acknowledged. Firstly, the performance on fine-grained MF values can vary. Although the measurement on some culturally sensitive values like "loyalty" is relatively reliable, others like "authority" and "sanctity" still miss significant cultural nuances.

Second, fine-tuning LLMs might backfire. Fine-tuning with limited local-language data may degrades model performance. While training with local language annotated data is commonly used to incorporate cultural specificity and enhance transfer learning, this approach appears less effective for LLMs in MF measurements task when data volume is small. We fine-tuned Llama3.1-8b using 20 batches of CCV data with the same strategy used in XLM-T. Each batch containing 50 MF labels evenly distributed across five classes. As shown in Figure 2, the initial drop in model performance struggles to recover within the available data for "care", "fairness", and "sanctity" values.

In addition, fine-tuning with English-annotated data may introduce extra cultural bias to LLMs. We replicated the same finetuning strategy to an Italian dataset "moralConvITA" (Stranisci et al. 2021), and observed similar significant performance drops in culturally-nuanced values like "loyalty" and "sanctity" (see Appendix). This signals the risk of cultural bias in English-centric LLMs. Finetuning such LLMs with English annotation data may exacerbate their bias against non-English languages, resulting in more cultural information loss in the measurement task, which echoes concerns about LLMs being shaped by values associated with WEIRD populations (Atari et al. 2023b). These findings highlight the need for culturally-sensitive fine-tuning approaches that preserve cross-cultural moral nuances.

Third, we note that our experiments were conducted in a single non-English language, which may limit the generalizability of our findings. Nevertheless, we hope this work can serve as a foundation for future research that extends to a broader range of non-English languages on MF measurement, particularly low-resource ones, where such methodological advancements would be especially valuable.

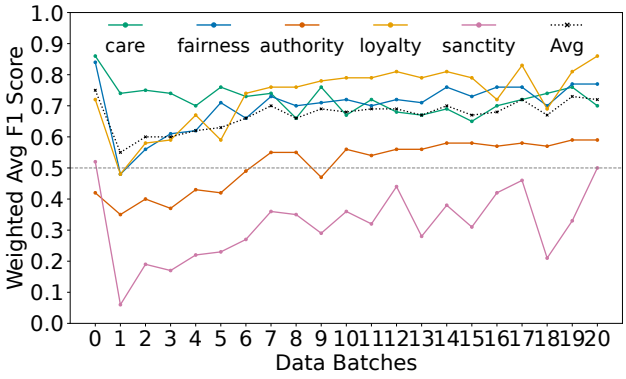

Figure 2: Accumulated fine-tuning Llama3.1-8b with local language annotated data from the Chinese CV dataset. Each batch includes 50 items evenly distributed across five classes.

Finally, we highlight the essential role of human validation in measuring cross-language MFs with LLMs. Given the limitations discussed, human validation is strongly recommended to ensure the reliability of LLM-based cross-language MF measurements. LLMs can effectively assist in this process by providing rationales alongside their classification outputs, supporting human evaluators and facilitating a more efficient result assessment.

Conclusions

This study examined four computational approaches—machine translation, dictionary, multilingual encoder-only models and decoder-only LLMs—for automatically measuring moral foundation values in non-English texts. It uses Chinese as a case study and leverages established English resources for this cross-language deductive coding task.

It first highlights the limitations of dictionary and machine translation approaches: while local language dictionaries can support lexicon-level analysis, they often lack the depth needed for complex semantic assessments. Notably, advanced English-based tools applied to machine-translated data outperform local lexicon-based methods, which underscores the limitations of lexicon approaches in this task. The study then explores the potential of transfer learning in language models, showing that both multilingual encoder-only models and LLMs demonstrate strong performance, with LLMs performing better and being more data-efficient. It is recommended to select approaches based on the availability of local-language annotated data. When sufficient data is available, a smaller multilingual language model generally yields satisfactory results. Otherwise, LLMs can serve as a reliable tool.

We recommend the following steps for applying LLMs to cross-language MF measurement: (1) start with multilingual LLMs and use culturally specific prompt engineering; (2) adopt a binary classification approach for each moral

foundation; (3) fine-tune models using available English annotations alongside carefully translated local-language data, while curating out English-centric value cases; and (4) incorporate human validation, especially for culturally distinctive values.

Findings in this paper provide valuable insights for cross-cultural MF research and shed light on future applications of LLM-assisted deductive coding in multilingual tasks.

Acknowledgements

The authors would like to thank Ralph Schroeder and Mohsen Molsleh for their feedback, and the Oxford Internet Institute for computing support.

References

Abdulhai, M.; Serapio-Garcia, G.; Crepy, C.; Valter, D.; Canny, J.; and Jaques, N. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17737–17752.

Agarwal, U.; Tanmay, K.; Khandelwal, A.; and Choudhury, M. 2024. Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language We Prompt Them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 6330–6340.

Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; Ramesh, K.; Jain, P.; Nambi, A.; Ganu, T.; Segal, S.; Ahmed, M.; et al. 2023. MEGA: Multilingual Evaluation of Generative AI. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Alper, S.; Bayrak, F.; Us, E. Ö.; and Yilmaz, O. 2020. Do changes in threat salience predict the moral content of sermons? The case of Friday Khutbas in Turkey. *European Journal of Social Psychology*, 50(3): 662–672.

Amin, A. B.; Bednarczyk, R. A.; Ray, C. E.; Melchiori, K. J.; Graham, J.; Huntsinger, J. R.; and Omer, S. B. 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1(12): 873–880.

Amin, M. M.; Cambria, E.; and Schuller, B. W. 2023. Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT. *IEEE Intelligent Systems*, 38(2): 15–23.

Araque, O.; Gatti, L.; and Kalimeri, K. 2020. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191: 105184.

Artetxe, M.; Labaka, G.; and Agirre, E. 2020. Translation Artifacts in Cross-lingual Transfer Learning. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7674–7684. Online: Association for Computational Linguistics.

Atari, M.; Haidt, J.; Graham, J.; Koleva, S.; Stevens, S. T.; and Dehghani, M. 2023a. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

Atari, M.; Xue, M. J.; Park, P. S.; Blasi, D. E.; and Henrich, J. 2023b. Which Humans? PsyArXiv:10.31234/osf.io/5b26t.

Atran, S. 2007. Religion, suicide, terrorism, and the moral foundation of the world. In *Social Brain Matters*, 101–117. Brill.

Barbieri, F.; Anke, L. E.; and Camacho-Collados, J. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 258–266.

Barriere, V.; and Balahur, A. 2020. Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, 266–271.

Brady, W. J.; Crockett, M. J.; and Van Bavel, J. J. 2020. The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 15(4): 978–1010.

Carvalho, F.; Okuno, H. Y.; Baroni, L.; and Guedes, G. 2020. A brazilian portuguese moral foundations dictionary for fake news classification. In *2020 39th International Conference of the Chilean Computer Science Society (SCCC)*, 1–5. IEEE.

Cheng, C. Y.; and Zhang, W. 2023. C-MFD 2.0: Developing a Chinese Moral Foundation Dictionary. *Computational Communication Research*, 5(2): 1.

Clifford, S.; Iyengar, V.; Cabeza, R.; and Sinnott-Armstrong, W. 2015. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4): 1178–1198.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzman, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

Dorothy, D.; Keiko, N.; and Vladimir, K. 2019. Lost in machine translation: Contextual linguistic uncertainty. *Bulletin of Volgograd State University. Series 2: Linguistics*, 18(4): 129–144.

Effron, D. A.; and Helgason, B. A. 2022. The moral psychology of misinformation: Why we excuse dishonesty in a post-truth world. *Current opinion in Psychology*, 47: 101375.

Frimer, J. A.; Boghrati, R.; Haidt, J.; Graham, J.; and Dehgani, M. 2019. Moral foundations dictionary for linguistic analyses 2.0. https://osf.io/ezn37/. Unpublished manuscript.

Gantman, A. P.; and Van Bavel, J. J. 2014. The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, 132(1): 22–29.

Gantman, A. P.; and Van Bavel, J. J. 2016. See for yourself: Perception is attuned to morality. *Trends in Cognitive Sciences*, 20(2): 76–77.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029.

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Haidt, J. 2012. The righteous mind: Why good people are divided by politics and religion. *Pantheon*.

Haidt, J.; and Graham, J. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1): 98–116.

Hoover, J.; Portillo-Wightman, G.; Yeh, L.; Havaldar, S.; Davani, A. M.; Lin, Y.; Kennedy, B.; Atari, M.; Kamel, Z.; Mendlen, M.; et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8): 1057–1071.

Hopp, F. R.; Fisher, J. T.; Cornell, D.; Huskey, R.; and Weber, R. 2021. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53: 232–246.

Hopp, F. R.; and Weber, R. 2021. Reflections on extracting moral foundations from media content. *Communication monographs*, 88(3): 371–379.

Hämmerl, K.; Deiseroth, B.; Schramowski, P.; Libovický, J.; Rothkopf, C. A.; Fraser, A.; and Kersting, K. 2023. Speaking Multiple Languages Affects the Moral Bias of Language Models. In *ACL (Findings)*, 2137–2156.

Iurino, K.; and Saucier, G. 2020. Testing measurement invariance of the Moral Foundations Questionnaire across 27 countries. *Assessment*, 27(2): 365–372.

Ji, J.; Chen, Y.; Jin, M.; Xu, W.; Hua, W.; and Zhang, Y. 2024. MoralBench: Moral Evaluation of LLMs. *CoRR*.

Jin, L.; He, J.; May, J.; and Ma, X. 2023. Challenges in Context-Aware Neural Machine Translation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jing, E.; and Ahn, Y.-Y. 2021. Characterizing partisan political narrative frameworks about COVID-19 on Twitter. *EPJ data science*, 10(1): 53.

Kertzer, J. D.; Powers, K. E.; Rathbun, B. C.; and Iyer, R. 2014. Moral support: How moral values shape foreign policy attitudes. *The Journal of Politics*, 76(3): 825–840.

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; Vidgen, B.; and Hale, S. A. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. arXiv:2404.16019.

Kivikangas, J. M.; Fernández-Castilla, B.; Järvelä, S.; Ravaja, N.; and Lönnqvist, J.-E. 2021. Moral foundations and political orientation: Systematic review and meta-analysis. *Psychological bulletin*, 147(1): 55.

Koleva, S. P.; Graham, J.; Iyer, R.; Ditto, P. H.; and Haidt, J. 2012. Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of research in personality*, 46(2): 184–194.

Kwak, H.; An, J.; Jing, E.; and Ahn, Y.-Y. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7: e644.

Lan, A. G. J.; and Paraboni, I. 2022. Text-and author-dependent moral foundations classification. *New Review of Hypermedia and Multimedia*, 28(1-2): 18–38.

Li, Z.; Shi, Y.; Liu, Z.; Yang, F.; Payani, A.; Liu, N.; and Du, M. 2024. Quantifying Multilingual Performance of Large Language Models Across Languages. arXiv:2404.11553.

Liscio, E.; Dondera, A. E.; Geadau, A.; Jonker, C. M.; and Murukannaiah, P. K. 2022. Cross-domain classification of moral values. In *2022 Findings of the Association for Computational Linguistics: NAACL 2022*, 2727–2745. Association for Computational Linguistics (ACL).

Liu, P.; Zhang, S.; Yu, D.; and Bo, L. 2022. CoreValue: Chinese Core Value-Behavior Frame and Knowledge Base for Value Computing). In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, 417–430.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Markowitz, E. M.; and Shariff, A. F. 2012. Climate change and moral judgement. *Nature climate change*, 2(4): 243–247.

Matsuo, A.; Sasahara, K.; Taguchi, Y.; and Karasawa, M. 2019. Development and validation of the Japanese moral foundations dictionary. *PloS one*, 14(3): e0213343.

Meta. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.

Mokhberian, N.; Abeliuk, A.; Cummings, P.; and Lerman, K. 2020. Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, 206–219. Springer.

Nguyen, T. D.; Chen, Z.; Carroll, N. G.; Tran, A.; Klein, C.; and Xie, L. 2024. Measuring Moral Dimensions in Social Media with Mformer. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1134–1147.

Nilsson, A.; and Erlandsson, A. 2015. The Moral Foundations taxonomy: Structural validity and relation to political ideology in Sweden. *Personality and Individual Differences*, 76: 28–32.

Nussio, E. 2023. How Moral Beliefs Influence Collective Violence. Evidence From Lynching in Mexico. *Comparative Political Studies*, 00104140231223747.

Peng, S.; Liu, C.; Deng, Y.; and Yu, D. 2021. Morality Between the Lines: Research on Identification of Chinese Moral Sentence). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 537–548.

Preniqi, V.; Ghinassi, I.; Ive, J.; Saitis, C.; and Kalimeri, K. 2024. MoralBERT: A Fine-Tuned Language Model for Capturing Moral Values in Social Discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, 433–442.

Princeton, U. 2010. About WordNet. https://wordnet.princeton.edu/. Online resource.

Ramesh, A.; Parthasarathy, V. B.; Haque, R.; and Way, A. 2021. Comparing statistical and neural machine translation performance on hindi-to-tamil and english-to-tamil. *Digital*, 1(2): 86–102.

Ranathunga, S.; Lee, E.-S. A.; Prifti Skenduli, M.; Shekhar, R.; Alam, M.; and Kaur, R. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11): 1–37.

Rathje, S.; Mirea, D.-M.; Sucholutsky, I.; Marjieh, R.; Robertson, C. E.; and Van Bavel, J. J. 2024. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34): e2308950121.

Röttger, P.; Nozza, D.; Bianchi, F.; and Hovy, D. 2022. Data-Efficient Strategies for Expanding Hate Speech Detection into Under-Resourced Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5674–5691.

Saunders, D. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75: 351–424.

Schuster, S.; Gupta, S.; Shah, R.; and Lewis, M. 2019. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 3795–3805. Minneapolis, Minnesota: Association for Computational Linguistics.

Shen, L.; Tan, W.; Chen, S.; Chen, Y.; Zhang, J.; Xu, H.; Zheng, B.; Koehn, P.; and Khashabi, D. 2024. The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts. arXiv:2401.13136.

Stasimioti, M.; Sosoni, V.; Kermanidis, K. L.; and Mouratidis, D. 2020. Machine Translation Quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. In *Proceedings of the 22nd annual conference of the European Association for Machine Translation*, 441–450.

Stranisci, M.; De Leonardis, M.; Bosco, C.; and Patti, V. 2021. The expression of moral values in the twitter debate: a corpus of conversations. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2): 113–132.

Tamborini, R.; Hofer, M.; Prabhu, S.; Grall, C.; Novotny, E. R.; Hahn, L.; and Klebig, B. 2020. The impact of terrorist attack news on moral intuitions and outgroup prejudice. In *Media, Terrorism and Society*, 66–90. Routledge.

Trager, J.; Ziabari, A. S.; Davani, A. M.; Golazizian, P.; Karimi-Malekabadi, F.; Omrani, A.; Li, Z.; Kennedy, B.; Reimer, N. K.; Reyes, M.; Cheng, K.; Wei, M.; Merrifield, C.; Khosravi, A.; Alvarez, E.; and Dehghani, M. 2022. The Moral Foundations Reddit Corpus. arXiv:2208.05545.

Troiano, E.; Klinger, R.; and Padó, S. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4340–4354.

Yilmaz, O.; Harma, M.; Bahçekapili, H. G.; and Cesur, S. 2016. Validation of the Moral Foundations Questionnaire in Turkey and its relation to cultural schemas of individualism and collectivism. *Personality and Individual Differences*, 99: 149–154.

Yu, L.; Leng, Y.; Huang, Y.; Wu, S.; Liu, H.; Ji, X.; Zhao, J.; Song, J.; Cui, T.; Cheng, X.; et al. 2024. CMoralEval: A Moral Evaluation Benchmark for Chinese Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, 11817–11837.

Zhang, X.; Li, S.; Hauer, B.; Shi, N.; and Kondrak, G. 2023. Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1): 237–291.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this work evaluates various computational methods for measuring moral foundations in Chinese text and highlights the promising performance of large language model-assisted approaches. These methods significantly enhance automatic cross-language moral foundation measurements, particularly in comparative studies within computational social science. The findings are descriptive and do not violate the above.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Related Work section.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes

(f) Did you discuss any potential negative societal impacts of your work? Yes. The Findings in this paper are useful in characterizing moral sentiment on social media in non-English text. We believe they do not pose any direct negative societal impacts.

(g) Did you discuss any potential misuse of your work? Yes, see the Discussion and Limitations section.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see the Discussion and Limitations section.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Yes, see the Related Work section.

(b) Have you provided justifications for all theoretical results? Yes, see the Discussion and Limitations section

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes

(e) Did you address potential biases or limitations in your theoretical framework? Yes

(f) Have you related your theoretical results to the existing literature in social science? Yes, moral foundation theory is a prominent theory in social psychology, see the Introduction.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, the results may provide valuable insights for future cross-cultural comparative studies on moral foundations in non-English languages. See the Discussion and Limitations section.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. The main paper and supplemental material describe datasets and training instructions in detail. Upon publication, these will be released publicly.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see the Methods section.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see the Methods section.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the Results section.

(f) Do you discuss what is "the cost" of mis-classification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets,

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? Yes, and cited the original sources of the datasets.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, our code including fine-tuning details can be accessed from GitHub.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, the datasets are public and we also have obtained consent from all original creators to use the datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, our datasets do not contain personally identifiable information or offensive content.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects,

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Additional Information of Models and Datasets

Figure 3 shows the curated mapping scheme for CCV dataset.

We provide additional details on model training and fine-tuning beyond those reported in the method section. Python scripts for replication are available on GitHub. All models were trained and fine-tuned on identical English annotation datasets: moral foundation Reddit corpus (MFRC), moral foundation Twitter corpus (MFTC), and extended moral foundation dictionary news (eMFD) corpus. Computations

Chinese Core Value			Moral Foundation Values	Virtue	Vice	All
category	subcategory	action_category				
文明	思想文明	宣传/学习知识/精神	sanctity/degradation	215	24	239
	公共文明	爱护公物	care/harm	9	62	71
		爱护环境	care/harm	72	40	112
		遵守公共秩序	authority/subversion	21	302	323
		积极参与社会治理活动	authority/subversion	46	0	46
	言语文明	用语礼貌	care/harm	11	40	51
	仪表文明	个人卫生	sanctity/degradation	0	8	8
		穿着服饰	sanctity/degradation	3	17	20
公正	思想公正	宣传公正言论	fairness/cheating	16		16
	机会公正	办事公正	fairness/cheating	180	147	327
		廉洁奉公	fairness/cheating	17	20	37
平等	思想平等	宣传平等言论	fairness/cheating	28	4	32
	人格平等	性别平等	fairness/cheating	39	102	141
		种族平等	fairness/cheating	17	55	72
		地域平等	fairness/cheating	1	44	45
		其他（人人平等）	fairness/cheating	94	130	224
法治	知法懂法	宣传/学习法律条款/内容	authority/subversion	16	7	23
	守法用法	遵守法律	authority/subversion	55	1027	1082
		配合民警执行公务	authority/subversion	17	110	127
爱国	思想爱国	宣传/学习爱国言论/知识	loyalty/betrayal	110	16	126
		心系祖国	loyalty/betrayal	129	7	136
	以身作则	维护祖国统一	loyalty/betrayal	190	17	207
		投入祖国建设	loyalty/betrayal	81	0	81
敬业	热爱岗位	热爱本职工作	loyalty/betrayal	51	1	52
	忠于职守	做好本职工作	loyalty/betrayal	1704	325	2029
		克制欲望	sanctity/degradation	26	121	147
诚信	诚实待人	真实诚恳	fairness/cheating	112	384	496
		拾金不昧	fairness/cheating	114	2	116
	信守诺言	兑现承诺	fairness/cheating	203	25	228
	传播诚信	宣传诚信	fairness/cheating	6	0	6
友善	乐于助人	见义勇为	care/harm	522	0	522
		捐款捐物	care/harm	502	0	502
		互相帮助（朋友、陌生人）	care/harm	1249	7	1256
	宽厚待人	以和为贵	care/harm	111	1284	1395
		关心慰问	care/harm	309	5	314
		孝顺长辈	authority/subversion	175	19	194
		爱护幼小	care/harm	136	68	204
		爱护动物	care/harm	3	1	4

Figure 3: Mapping Scheme of Chinese Core Value Dataset

were performed on a single NVIDIA L40S GPU. Please see following links for models and datasets used in this paper.

MFRC — https://paperswithcode.com/dataset/mfrc

MFTC — https://osf.io/k5n7y/

eMFD — https://osf.io/preprints/psyarxiv/924gq_v1

MoralBert — https://huggingface.co/vjosap

Mformers — https://huggingface.co/joshnguyen

c2-roberta-base-finetuned-dianping-chinese — https://huggingface.co/liam168/c2-roberta-base-finetuned-dianping-chinese

XLM-T — https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base

XLM-T Task objective: Binary classification of five moral foundation values. Base model: `cardiffnlp/twitter-xlm-roberta-base` from Hugging Face. Hyperparameters: Learning rate: $2e-5$. Epochs: 3. Batch size: 16 (train & eval). Weight decay: 0.01. Warmup steps: 100.

Llama3.1 Task objective: Multiclass classification of five moral foundation values. Base model: `llama3.1-8b-instruct` from Hugging Face. Framework: Unsloth with PEFT via LoRA. Precision: LoRA applied with 4-bit quantization. We used the llama-3 chat template from Unsloth. Hyperparameters: Learning rate: $2e-5$. Epochs: 3. Batch size: 128 (train & eval). Weight decay: 0.01. Max Seq Length: 1024. LoRA configurations used default settings: r-16, targeting q, k, v, o, gate, up, and down-proj parameters. Random state was set to 3047. The English system prompt is shown below. Chinese and Italian prompts were translated accordingly and shared on the GitHub. Few-shot examples ($N = 3$) were purposefully sampled from the benchmarking dataset in the local language.

Prompt

You are a native Chinese speaker and social science annotator, your task is to label the moral foundation values expressed in the given Chinese documents.

Moral foundation values are the core values that underlie moral reasoning from the moral foundation theory. the five moral foundations are: care, fairness, loyalty, authority, and sanctity. And they refer to the following moral intuitions, each includes both vice/virtue pairs:

- care: related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies the virtues of kindness, gentleness, and nurturance.
- fairness: related to the evolutionary process of reciprocal altruism. It underlies the virtues of justice and rights.
- loyalty: our long history as tribal creatures able to form shifting coalitions. It is active anytime people feel that its "one for all and all for one." It underlies the virtues of patriotism and self-sacrifice for the group.
- authority: This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to prestigious authority figures and respect for traditions.
- sanctity: This foundation was shaped by the psychology of disgust and contamination. It underlies notions of striving to live in an elevated, less carnal, more noble, and more natural way (often present in religious narratives). This foundation underlies the widespread idea that the body is a temple that can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions). It underlies the virtues of self-discipline, self-improvement, naturalness, and spirituality.

you should follow the given principles to label the moral foundation values in the give documents:

1. identify the moral foundation value only from the 5 given ones.
2. if the document expresses more than 1 foundation value, label all prominent values, but in total should be equal or less than 3 values.
3. provide a brief rationale for the each labelling, which should be less than 20 words.
4. labels the value in English
5. rationales should be in the same language as the document
6. if the document does not express any of the 5 values, label it as 'none' and provide a brief rationale.
7. if the document can not be labelled into any of the 5 values, label it as 'unknown' and provide a brief rationale.
8. consider the Chinese cultural context of the document when labelling the values.

You MUST respond with a brief rationale within 15 words, and the labels. save in the dictionary format:

"rationale": "reasons to explain your decision", "labels": "you labels here"

Here are the given documents for your task:

Additional Qualitative Analysis

Figure 4 shows examples of mis-annotated texts in the CCV dataset. English translations were generated via Google Translate using the `deep-translator` Python package. MFormer shows results from machine-translated English text, while Llama shows results from original Chinese texts. We used the Llama3.1-8b-instruct model fine-tuned on English annotations and their Chinese translations.

Replicate LLMs Approach on Italian

We tested a similar LLM-based strategy on an Italian dataset "moralConvITA" (Stranisci et al. 2021), with results shown in Table 7. This benchmark comprises 1,724 unique Twitter posts discussing the immigration issue in Italy, labeled by native Italian speakers ($N = 8$) according to five moral

No	Chinese	English Translation	Human	MFormer	Llama
1	你啊你总是向着自己的小儿子	You, you always look towards your little son	fair	care	none
2	老人故意碰瓷儿。	The old man touched the porcelain on purpose.	fair	authority	fair
3	他本可颐养天年，却主动要求挂职农村，在岗位上苦干实干，用一颗坚定的爱民之心带领村民奔小康。	He could have taken care of his old age, but he took the initiative to work in rural areas, worked hard at his post, and led the villagers to a well-off society with a firm love for the people.	loyalty	care	loyalty
4	黄宝妹将“听党的话”作为自己的信念。	Huang Baomei regards "listening to the party's words" as her belief.	loyalty	authority	authority
5	伍庭光是香港新来港人仕服务基金副会长，爱国爱港，为维护香港繁荣稳定	Wu Tingguang is the vice president of the Hong Kong New Arrivals Service Fund. He loves the country and Hong Kong and strives to maintain the prosperity and stability of Hong Kong.	loyalty	loyalty	authority
6	母亲的病不断加重，李灿琳更是端屎端尿，照顾极为细心，就像小时候母亲照顾她一样。	As her mother's illness continued to worsen, Li Canlin was taking care of her very carefully, just like her mother took care of her when she was a child.	authority	care	care
7	广东一训练营教官对戒网少年实行殴打、禁止喝水等惩罚方式，造成少年肾衰竭。	Instructors at a training camp in Guangdong punished teenagers who quit using the Internet by beating them and depriving them of water, causing the teenager to suffer from kidney failure.	authority	care	sanctity
8	宝鸡一个农家小院里，90 多岁高龄的黄清海老人安逸地晒着太阳。该吃午饭了，张引乾端来特意给老母亲做的软面片，等老人吃完安顿好，张引乾才匆匆吃起午饭。	In a small farmyard in Baoji, Huang Qinghai, an old man in his 90s, was basking in the sun comfortably. It was time to have lunch. Zhang Yingqian brought the soft noodles specially made for his old mother. After the old man finished eating and settled down, Zhang Yingqian ate lunch in a hurry.	authority	care	none
9	孙艳华一直精心照顾着父亲，直到父亲去世	Sun Yanhua took good care of her father until his death	authority	care	care
10	对于婆婆的一切举动，孙斌云从不责骂和抱怨，只是一心一意地悉心照料。	Sun Binyun never scolded or complained about her mother-in-law's actions, but took good care of her wholeheartedly.	authority	care	care
11	张引大肆收受下级以奖金名义的贿赂，又滥用职权违规发放奖金，	Zhang Yin recklessly accepted bribes from his subordinates in the name of bonuses, and abused his power to issue bonuses in violation of regulations.	sanctity	fair	fair
12	他表示，学校要弘扬“苏兆征精神”——勇往直前的拼搏精神，实事求是的科学精神，淡泊名利的求实精神，真诚和善的合作精神。	He said that the school should carry forward the "Su Zhaozheng Spirit" - the courageous fighting spirit, the scientific spirit of seeking truth from facts, the realistic spirit of being indifferent to fame and fortune, and the sincere and kind spirit of cooperation.	sanctity	sanctity	loyalty
13	河北男子听信半仙	Hebei man listens to half-immortal	sanctity	authority	sanctity

Figure 4: Qualitative analysis of cultural nuances. *Chinese* column refers to the target Chinese text; *English Translation* column is the result of machine translation using Google Translate. *Human, MFormer and Llama* refer to the labels annotated by human annotators, MFormer, and Llama3.1-8b-instruct model.

foundations. To minimize annotation ambiguity, we sampled approximately 100 texts with distinct labels for each moral foundation.

We applied the same supervised instructive fine-tuning strategy used in the CCV dataset. First, we used base Llama3.1-8b-instruct model with both English and Italian prompts. Then, we fine-tuned the model with English annotated text-label pairs from Twitter, Reddit, and news corpora used on CCV. Finally, we performed data augmentation using Italian translations generated via Google Translate. We maintained similar default settings with learning rate, epochs, and batch size at $2e$-5, 1, and 16, respectively.

We observed similar results as shown in CCV with some noticeable distinctions. Finetuning with translated local language annotations shows strong potential in enhancing LLMs' performance on culturally distinct MF values in CCV. For example, Llama3.1-8b notably improves on "loyalty" when fine-tuned with English data machine-translated to Chinese ($F1 = 0.69$), compared to its performance with the original English training data ($F1 = 0.57$). Such strategy may serve as a practical and cost-effective way to improve model performance in cross-language measurements, especially in cases where mass-labeling of local-language data is infeasible due to the resource-intensive process of creating high-quality language-specific labeled datasets.

This strategy, however, showed moderate performance on the Italian dataset. This difference may be attributed to the smaller Italian benchmark sample size ($N = 492$ vs. approximately 1,500 in CCV), the unstructured nature of Twitter data (typically brief with limited context), and differences in domain and language complexity. The brevity and informality of Italian tweets likely posed greater challenges for accurate moral foundation inference.

	Auth	Care	Fair	Loya	Sanc	Acc	Cov	Fw	Fm
baseline	0.20	0.20	0.20	0.20	0.20	0.20	1.00	0.20	0.20
en×∅	0.53	**0.52**	**0.43**	**0.57**	0.26	**0.48**	0.75	**0.46**	**0.46**
it×∅	0.54	0.44	0.27	0.35	0.30	0.38	0.39	0.38	0.38
en × en	0.57	0.48	0.43	0.25	0.24	0.44	1.00	0.39	0.39
it × en	0.69	0.27	0.09	0.42	0.4	0.42	**0.94**	0.37	0.37
en × it	0.22	0.49	0.32	0.38	0.16	0.34	0.76	0.31	0.31
it × it	**0.72**	0.36	0.09	0.51	**0.35**	0.46	0.82	0.41	0.41
en ×(en+it)	0.48	0.47	0.37	0.04	0.02	0.37	1.00	0.27	0.27
it ×(en+it)	0.43	0.38	0.16	0.64	0.06	0.39	**0.94**	0.34	0.33

Table 7: Llama3.1-8b model for moral foundation measurements on sampled MoralCovnITA dataset. The first language refers to the language of prompt while the languages after ×refer to the language(s) of fine-tuning datasets. ∅ denotes no fine-tuning. The best performing method is in **bold**. Column labels are consistent with Table 2.

Then we further qualitative assessed the performance on some moral foundation categories and revealed the cultural misalignment limitation in the finetuning process. Finetuning with English annotations sometimes produced negative effects. Performance on culturally sensitive foundations such as "loyalty" and "sanctity" declined significantly after fine-tuning with English and translated Italian data. This signals a concerning cultural misalignment in current LLMs — fine-tuning with English-centric moral foundations may exacerbate inherent cultural biases. Our qualitative reading of some mislabeled Italian texts supports this finding. For example, Italian understanding of sanctity, influenced by Catholic values and historical context, differs significantly from American conceptions, resulting in poor performance after fine-tuning. Future research should further investigate fine-tuning techniques that mitigate cultural misalignment.

Moreover, we acknowledge the challenges involved in human annotation of moral foundations. Both CCV and MoralConvITA datasets have limitations, with annotator disagreements being common — even the same annotator may label identical text differently at different times. Although training can reduce inconsistency, maintaining quality control remains difficult at scale. Both datasets relied on expert coders, which may introduce labeling biases. Additionally, MoralConvITA is subject to topic and text form biases, as it focuses on immigration during a specific period and consists of Twitter posts. In contrast, the CCV dataset covers a broader range of topics and primarily includes posts from news articles. Consequently, a general performance drop on the Italian dataset is expected. These challenges underscore the importance of creating high-quality benchmarking datasets for moral foundation measurement. Future research should consider employing demographically representative crowd workers to better capture the distribution of cultural values in local contexts (Hopp and Weber 2021).

In conclusion, LLMs offer a promising approach for measuring moral foundations in non-English texts when limited local language annotations are available. With careful prompting and few-shot learning, LLMs deliver moderate performance compared to human annotation. It should be noted that fine-tuning with English annotations though may improve the performance on some moral foundation values, it can introduce additional biases to cultural-sensitive ones, therefore diminish the model performance.