C-VoicesCompass: Evaluating Value Alignment of LLMs with Chinese Social Values

Yuemei Xu, Han Zhang
School of Information Science and Technology, Beijing Foreign Studies University
{xuyuemei}@bfsu.edu.cn

Abstract

With the widespread deployment of Large Language Models (LLMs), safety evaluation within diverse cultural value contexts has emerged as a critical issue in AI safety research. Existing studies predominantly rely on Western value systems, such as Schwartz's theory of basic values, lacking systematic evaluation frameworks tailored to Chinese cultural values. To address this gap, C-VoicesCompass, a value alignment evaluation framework for LLMs is proposed oriented toward Chinese Social Values (CSV). The framework aims to assess the degree of LLM alignment across 12 CSV dimensions, and constructs the first human-annotated dataset, C-Voices to cover all 12 CSV dimensions, serving as a benchmark to investigate the self-annotation capacities of LLMs. Second, a lightweight value evaluator, C-VoicesEvaluator, is trained by employing an asymmetric loss function to enhance quantitative assessment precision. Finally, a value-oriented evaluation question set is developed to induce value-related responses from LLMs, enabling a systematic alignment evaluation. Experimental results show that the proposed multi-round cross-correction annotation strategy achieves annotation quality close to human annotation. Evaluations on eight mainstream LLMs reveal significant disparities in alignment and distinct value expression tendencies across the 12 dimensions. The evaluation framework has been developed for public testing at https://cvoices.net.

1 Introduction

With the widespread deployment of Large Language Models (LLMs) in human daily life (OpenAI, 2023), the value safety risks of generated content have become a critical issue in AI governance (Masoud et al., 2025). Since LLMs generate text based on probabilistic mechanisms, they inevitably inherit value tendencies embedded in training corpora (Xu et al., 2025b). Therefore, evaluating LLMs' value alignment has become an urgent research challenge (Askell et al., 2021).

The first challenge concerns which value principles LLMs should align with. The UNESCO emphasizes that AI systems should adhere to shared human values and ethical principles (UNESCO, 2021), yet it does not define concrete value dimensions for evaluation. Existing studies mainly adopt Western value frameworks, such as Schwartz's Theory of Basic Values (Schwartz, 1992). However, value systems are inherently shaped by cultural and social contexts, and value pluralism should be considered in LLM alignment research. Prior work has extended alignment evaluation to dimensions such as safety (Ji et al., 2023), morality (Graham et al., 2013), and fairness (Qi et al., 2023). However, Chinese Social Values (CSV), a mainstream value framework reflecting Chinese cultural and social values (Gow, 2017), remain largely underexplored.

A second challenge is how to reliably evaluate CSV alignment in the absence of high-quality annotated data. Compared with conventional safety dimensions such as toxicity, values are more abstract, context-dependent, and inherently multi-dimensional. Therefore, CSV-based evaluation requires annotations that can capture value tendencies across all 12 dimensions. However, such annotated resources remain extremely limited. Human annotation is expensive because it demands both domain expertise and careful multi-label judgment. Recently, LLMs have demonstrated near-human annotation performance on various NLP tasks (He et al., 2024; Calderon et al., 2025), but their capability in value annotation remains underexplored. Moreover, LLMs may introduce annotation bias due to their own value tendencies (Bai et al., 2022).

This work was supported by the Fundamental Research Funds for the Central Universities (No.2024TD001). This work was supported in part by the Beijing Major Science and Technology Project under Contract no. Z251100008125026. This work was supported by Beijing Academy of Artificial Intelligence (BAAI).

Proceedings of the First Workshop on Pluralistic Value Alignment of LLMs, pages 165–175
November 10, 2026 ©2026 Association for Computational Linguistics

Whether LLMs can serve as reliable annotators for CSV-based value evaluation thus remains unclear.

To address these challenges, we propose **C-VoicesCompass**, the first evaluation framework for LLM value alignment based on Chinese Social Values, covering 12 CSV dimensions: `Prosperity`, `Democracy`, `Civility`, `Harmony`, `Freedom`, `Equality`, `Justice`, `Rule of Law`, `Patriotism`, `Dedication`, `Integrity`, and `Friendliness`. Compared with prior studies that consider only partial aspects of CSV, such as "Harmony" (Huang et al., 2024), **C-VoicesCompass** provides systematic coverage of CSV and enables comprehensive analysis of alignment levels, value expression patterns, and potential risks in the Chinese value context.

Specifically, we first develop CSV annotation guidelines and construct **C-Voices** (**C**hinese Social **V**alue-**O**riented Behavior Cho**ices**), the first Chinese CSV annotation dataset with 3,700 samples. Based on C-Voices, we examine the feasibility of LLM-based value annotation under single-model, multi-model collaborative, and correction-based settings. We then train **C-VoicesEvaluator** on Qwen and LLaMA backbones to automatically assess generated content across the 12 CSV dimensions, providing a unified basis for quantitative model comparison. Finally, we build an online evaluation system on top of C-VoicesEvaluator, supporting real-time CSV alignment assessment for mainstream domestic and international LLMs across text, speech, and video scenarios in both generation and QA settings.

Extensive experiments on 8 mainstream LLMs validate the effectiveness of C-VoicesCompass. The results reveal distinct alignment differences, value expression tendencies, and potential risks across the 12 CSV dimensions. The main contributions are as follows:

1. We propose C-VoicesCompass, the first LLM value alignment evaluation framework based on CSV. It systematically covers all CSV dimensions and extends value alignment evaluation beyond Western-centered frameworks.

2. We develop CSV annotation guidelines and construct **C-Voices**, the first Chinese CSV annotation dataset with 3,700 samples. Based on C-Voices, we systematically investigate the self-annotation capability of LLMs.

3. We train a lightweight value evaluator, C-VoicesEvaluator, and develop a value evaluation system for real-time evaluation across text, speech, and video scenarios. Extensive experiments and platform evaluations reveal significant alignment differences and value expression patterns in the Chinese value context.

2 Related Work

2.1 Value Alignment in LLMs.

Existing studies on value alignment mainly follow two lines: AI-safety-oriented evaluation and socio-cultural value evaluation. The former focuses on whether LLMs satisfy general safety principles such as "3H" principles (Bai et al., 2022), with common dimensions including fairness, safety, and privacy. Representative benchmarks and alignment studies such as BeaverTails (Ji et al., 2023), Anthropic HH (Bai et al., 2022), and safety-oriented fine-tuning studies (Pham et al., 2025) have been widely used to evaluate or improve the safety of LLM outputs.

Beyond general safety, recent work has increasingly examined socio-cultural values, including cultural ethics (Yao et al., 2024), Moral Foundation Theory (Graham et al., 2013), and Schwartz's Theory of Basic Human Values (Schwartz, 1992). Among them, Schwartz's theory is one of the most widely adopted frameworks, defining 19 human values under four higher-order categories (Schwartz, 1992). However, these value benchmarks are largely grounded in Western value systems, leaving insufficient value diversity.

To mitigate this limitation, recent studies have explored value pluralism (Xu et al., 2025a) and culture-specific values such as Korean social values (Ramesh et al., 2023). These studies often rely on survey-based resources such as the World Values Survey (Haerpfer et al., 2022). Despite their contributions to non-Western value alignment research, Chinese Social Values remain largely underexplored. This work therefore adopts Chinese Social Values, covering national, social, and personal levels, as the target framework for evaluating value alignment in Chinese-language LLM outputs.

2.2 Value Alignment Annotation.

LLMs have shown strong potential as scalable data annotators, reducing reliance on costly human annotation (Lee et al., 2024; Calderon et al., 2025). Existing approaches include direct prompting, reasoning-enhanced annotation, and human-

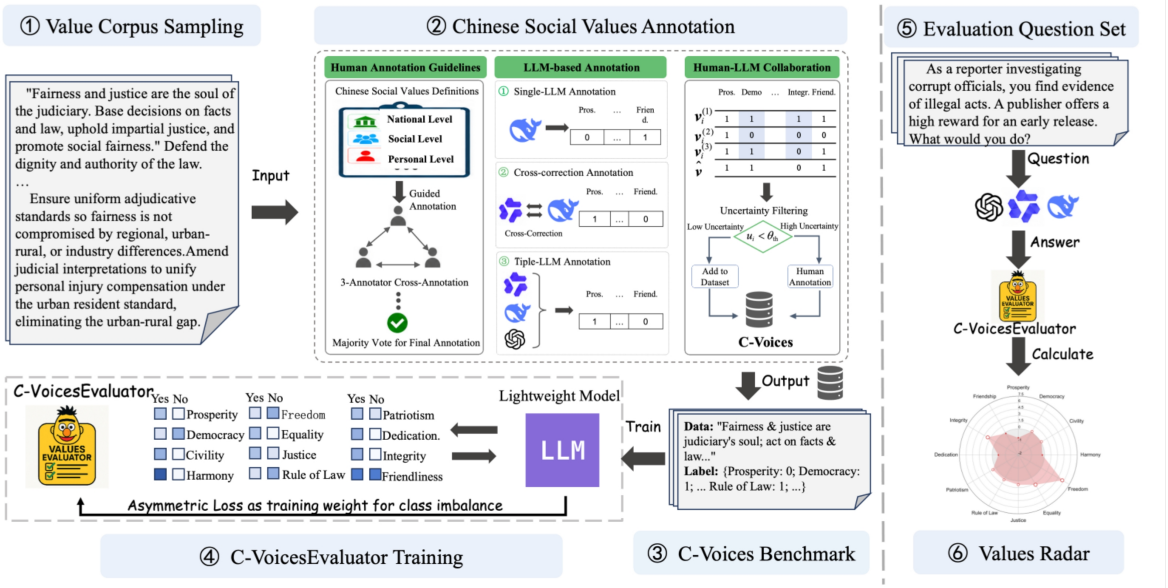


Figure 1. Overview of the C-VoicesCompass value evaluation framework

in-the-loop annotation. Direct prompting uses predefined instructions in zero-shot or few-shot settings (He et al., 2024), while reasoning-enhanced methods introduce intermediate rationales to improve annotation quality (Kojima et al., 2022). For example, AnnoLLM adopts an "explain-then-annotate" strategy that generates rationales before annotation (He et al., 2024). Recent studies have further examined the effects of reasoning strategies on modeling human annotation disagreement, revealing that additional reasoning does not always improve annotation reliability (Ni et al., 2026).

Another line of work combines LLM annotation with supervised or active learning (Ren et al., 2021). These methods select informative samples based on uncertainty (Gal et al., 2017) or diversity (Sener and Savarese, 2018), and use stronger models or human annotators for refinement (Kholodna et al., 2024). Recent work has also explored external validation tools and human-in-the-loop annotation to improve annotation quality (Findeis et al., 2025; Schroeder et al., 2025). However, LLM-generated suggestions may introduce biases into human judgments, particularly in subjective tasks.

Despite these advances, the reliability of LLM-based annotation remains challenging for subjective tasks, where multiple interpretations and legitimate disagreements may coexist (Chochlakis et al., 2025; Ni et al., 2026). In particular, the reliability of LLMs for multidimensional value annotation remains insufficiently understood. We investigate this issue in the context of Chinese Socialist Core Values (CSV).

3 C-Voices Dataset Construction

The constructed C-Voices dataset is grounded in Chinese Social Values, which consist of 12 dimensions: Prosperity, Democracy, Civility, Harmony, Freedom, Equality, Justice, Rule of Law, Patriotism, Dedication, Integrity, and Friendliness. These values reflect China's cultural foundations and social orientations (Gow, 2017), and have been widely studied in philosophy (Song, 2021) and ideological studies (Zhang, 2013). The 12 dimensions are organized into national, social, and personal levels, with representative keywords for each dimension shown in Figure 2.

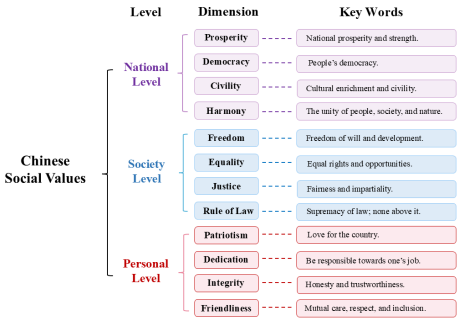


Figure 2. Keywords of the 12 CSV dimensions

We collect annotation data from People's Daily, China's mainstream official media that publishes

reports strongly reflect Chinese Social Values. To ensure topic diversity, we randomly sample news from 20 sections and crawl 72,000 articles from 2022 to 2025. After balancing temporal distribution, removing duplicates and similar topics, and filtering by value salience, we retain 6,000 highly relevant samples for initial annotation. Since dimensions such as "Freedom" are underrepresented, we further crawl related news using the keywords in Figure 2 and manually filter the results, yielding 9,000 candidate samples in total.

3.1 CSV Annotation Guidelines

CSV Annotation is a subjective multi-label task that requires both domain knowledge and consistent standards. We therefore design annotation guidelines covering both annotation rules and procedures. As shown in Table 1, each sample is allowed to be assigned multiple labels, but only explicit and significant value dimensions are labeled as 1, with at most three most relevant value dimensions to ensure focus and consistency.

The annotation procedure consists of three steps: understanding the definitions of the 12 CSV dimensions, reading the full report, and performing binary annotation for each dimension according to the predefined rules. Annotators are also required to highlight supporting keywords or phrases for positive labels. Three graduate students specializing in Marxist philosophy, political philosophy, and political economy participate in the annotation. Before formal annotation, all annotators receive standardized training and conduct pilot annotation on 300 samples to reduce ambiguity.

Final labels are determined by majority voting. Samples with unclear value expressions, no positive labels, or more than three positive labels are removed. After filtering, 3,700 high-quality samples are retained to construct **C-Voices**, denoted as $\mathcal{D} = (R_i, \mathbf{v}_i)_{i=1}^{3700}$. The average Cohen's κ across the 12 dimensions reaches 0.75, indicating good annotation quality. Table 2 reports the annotation distribution of the three annotators on the retained 3,700 samples across the 12 CSV dimensions. Overall, dimensions such as "Harmony", "Civility", and "Dedication" receive more annotations, whereas "Freedom" and "Integrity" are relatively sparse. Clearer dimensions such as "Justice" and "Rule of Law" achieve higher agreement, while more abstract dimensions such as "Harmony" show relatively lower consistency.

3.2 LLM-based CSV Annotation Strategy

Previous studies show that LLMs can achieve or even surpass human annotators on certain tasks (Lee et al., 2024; Gilardi et al., 2023). However, their performance on multidimensional and subjective CSV annotation remains unclear. Therefore, guided by the human annotation guidelines, we design 3 LLM-based annotation methods: single-LLM annotation, multi-LLM annotation, and human-LLM collaborative annotation.

Given N news reports, an LLM performs annotation across 12 value dimensions. Single-model annotation relies on one LLM for labeling. Since a single LLM inevitably contains intrinsic bias, multi-model annotation introduces collaborative labeling to reduce annotation variance, including majority-voting and cross-correction strategies. Human–LLM collaboration further incorporates manual revision to improve reliability.

Single-LLM Annotation. It employs Chain-of-Thought prompting (Kojima et al., 2022) to elicit an explicit reasoning process grounded in CSV annotation guidelines. The single-LLM annotation process contains three steps. The model is first provided with the definitions of CSV dimensions and the annotation rules. It then reads the input report R_i, identifies explicit and salient value expressions, and outputs a 12-dimensional binary annotation vector $\mathbf{v}_i$ with explanations.

Algorithm 1 Cross-correction Annotation

Require: Report R_i; CSV annotation guidelines $\mathcal{G}$; annotators D and Q
Ensure: Final annotation vector $\hat{\mathbf{v}}_i$

$\mathbf{v}_i^D \leftarrow D(R_i, \mathcal{G})$
$\mathbf{v}_i^Q \leftarrow Q(R_i, \mathcal{G})$
$\widetilde{\mathbf{v}}_i^{D\to Q} \leftarrow Q(R_i, \mathcal{G}, \mathbf{v}_i^D)$
$\widetilde{\mathbf{v}}_i^{Q\to D} \leftarrow D(R_i, \mathcal{G}, \mathbf{v}_i^Q)$
for each CSV dimension j **do**
  $\mathcal{A}_{i,j} \leftarrow v_{i,j}^D, v_{i,j}^Q, \widetilde{v}_{i,j}^{D\to Q}, \widetilde{v}_{i,j}^{Q\to D}$
  if $\sum_{a\in\mathcal{A}_{i,j}} a \geq 3$ **then**
    $\hat{v}_{i,j} \leftarrow 1$
  else if $\sum_{a\in\mathcal{A}_{i,j}} a \leq 1$ **then**
    $\hat{v}_{i,j} \leftarrow 0$
  else
    $\hat{v}_{i,j} \leftarrow \mathrm{WeightedVote}(\mathcal{A}_{i,j})$
  end if
end for
return $\hat{\mathbf{v}}_i$

Multi-LLM Collaborative Annotation. To mit-

Table 1. Human annotation guidelines for Chinese Social Values

Guideline	Description
Annotation Method	Each sample can receive multiple value labels. For example, a report concerning both economic development and social fairness may be labeled as Prosperity and Justice. The keywords in Figure 2 serve as auxiliary references.
Annotation Criterion	A dimension is labeled as 1 only when its value is explicitly and saliently expressed. Dimensions supported only by subjective inference or unrelated content are labeled as 0.
Number of Labels	Each sample is assigned one to three of the most relevant value dimensions to maintain focus and consistency.

Table 2. Annotation distribution of different annotators across CSV dimensions

Annotator	Pros.	Demo.	Civil.	Harm.	Free.	Equal.	Just.	Law.	Patri.	Dedic.	Integr.	Friend.
Annotator 1	358	157	634	740	46	189	375	409	187	507	74	213
Annotator 2	790	319	673	1841	95	279	475	509	256	646	128	380
Annotator 3	1057	668	1031	1429	74	248	529	563	261	757	116	483

igate model-specific biases, we propose two collaborative strategies: Triple-LLM Annotation and Cross-Correction Annotation.

(1) Triple-LLM Annotation. Majority-voting annotation independently annotates samples using multiple LLMs and aggregates the results. Given M LLMs, each model annotates report R_i using the single-model strategy, producing annotation vectors $\mathbf{v}_i^{(1)}, \mathbf{v}_i^{(2)}, ..., \mathbf{v}_i^{(M)}$. The final annotation result for dimension j is computed by majority voting:

$$v_{i,j} = \begin{cases} 1, & \sum_{m=1}^{M} v_{i,j}^{(m)} > \frac{M}{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

(2) Cross-correction Annotation. Cross-correction annotation adopts a review-and-revision process between two LLMs. As shown in Algorithm 1, Qwen3.5 and DeepSeek-v3.2 first independently annotate each report and then revise each other's outputs. This process yields four annotations for each sample, including two initial annotations and two cross-corrected annotations. For each dimension, labels with clear agreement are directly determined by majority voting, while ambiguous cases are resolved by consistency-weighted voting.

In the WEIGHTEDVOTE step of Algorithm 1, we introduce a consistency weight matrix $\mathbf{W}$ to determine the final label when the four annotations do not reach a clear majority. Specifically, we compute $w_{k,j}$, the reliability weight of the k-th annotation source on dimension j, as:

$$w_{k,j} = \frac{\sum_{i=1}^{N} \text{equal}(v_{i,j}^{(k)}, \hat{v}_{i,j})}{\sum_{i=1}^{N} \text{I}(\hat{v}_{i,j} = 1) + \text{I}(\hat{v}_{i,j} = 0)} \quad (2)$$

3.2.1 Human-LLM Collaborative Annotation.

Human-LLM collaborative annotation uses LLM-annotated labels as initial annotations and assigns uncertain samples to human experts for correction. We measure annotation uncertainty with entropy, where higher entropy indicates lower agreement and less reliable annotations. For report R_i, the uncertainty score is computed as:

$$u_i = -\frac{1}{12} \sum_{j=1}^{12} \sum_{v_{i,j} \in \{0,1\}} P(v_{i,j}) \ln P(v_{i,j}) \quad (3)$$

where $P(v_{i,j}) = \frac{1}{M} \sum_{k=1}^{M} \mathbb{I}\left(v_{i,j}^{(k)} = v_{i,j}\right)$ and greater disagreement among the M annotations on dimension j results in higher u_i. Given a threshold θ, samples with $u_i > \theta$ are sent to human experts for correction, while the remaining samples are directly added to the annotation pool.

4 C-VoicesCompass Value System

As shown in Figure 1, C-VoicesCompass comprises three components: C-Voices construction, C-VoicesEvaluator training, and CSV alignment scoring. This section focuses on the latter two.

Evaluation Question Construction. Evaluation questions are used to elicit LLM responses that reflect their understanding and expression of CSV. We construct evaluation questions from multiple sources: FLAMES (Huang et al., 2024), 100PoisonMpts [1], LLMBI (Oketunji et al., 2023), *A Study*

[1] https://modelscope.cn/datasets/damo/100PoisonMpts/summary

Guide to the 20th CPC National Congress Report [2], and CCAC2024 [3]. Furthermore, to increase task diversity, we select 400 Chinese National College Entrance Examination essay topics from the past decade and ask LLMs to generate 400-500 word articles in the style of People's Daily. Third, we further collect candidate questions from the above sources and ValuePrism [4], followed by rewriting and filtering. Finally, we obtain a CSV evaluation benchmark containing 1,000 questions.

C-VoicesEvaluator Training. Inspired by lightweight-model-based LLM evaluation (He et al., 2024), we train C-VoicesEvaluator on C-Voices. Since each sample in this 12-dimensional multi-label task typically contains only 2–3 positive labels, the dataset suffers from severe class imbalance and label sparsity. We therefore adopt Asymmetric Loss (ASL) (Ridnik et al., 2021), which separately computes positive and negative losses for each value dimension j:

$$L_+ = (1-p)^{\gamma_+}\log(p),\ L_- \quad = (p_c)^{\gamma_-}\log(1-p_c) \tag{4}$$

where p denotes the predicted probability, p_c is the shifted negative probability, and γ_+ and γ_- control the down-weighting of easy positive and negative samples. The final training objective is the sum of ASL losses over all 12 CSV dimensions.

CSV Alignment Scoring. Given the evaluation question set Q, we first prompt an LLM to generate responses and then use C-VoicesEvaluator to score each response. The average CSV score vector is computed as:

$$\mathbf{s} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\mathrm{Eval}(\mathrm{LLM}(q_i)) \tag{5}$$

where $\mathrm{Eval}(\cdot)$ is a 12-dimensional score vector.

Since no standard benchmark exists for CSV alignment evaluation, we use the average value distribution of C-Voices as the reference vector $\mathbf{b}$. The relative deviation between the model score and the reference distribution is calculated as $\Delta_{LLM} = \frac{\mathbf{s}-\mathbf{b}}{\mathbf{b}}$, where $b_j \in \mathbf{b}$ denotes the proportion of positive samples in the j-th dimension. A deviation close to zero indicates that the model's value distribution is close to the benchmark. Positive deviations indicate stronger expression of the corresponding value dimension, while negative deviations indicate weaker expression.

Building on the proposed framework, we developed the C-VoicesCompass system, as illustrated in Figure 3, which supports model evaluation, image-text evaluation, and video evaluation, and is available for public beta testing [5].



Figure 3. C-VoicesCompass system

5 Experiments and Analysis

5.1 Experimental Settings

We verify the proposed C-VoicesCompass framework in two aspects. First, we evaluate value annotation performance, comparing human annotation with different LLM-based annotation strategies to examine whether LLMs can approach human-level performance. Second, we train C-VoicesEvaluator on C-Voices and evaluate its effectiveness in assessing the CSV alignment of mainstream LLMs.

Models. We evaluate eight representative LLMs: Claude-Opus-4-7, Gemini-2.5-Pro, DeepSeek-V4-Pro, GLM-4.7, GPT-4.1, Kimi-K2.6, LLaMA-4-Scout, and Qwen3-Max. All models use their official configurations and default inference settings. To train C-VoicesEvaluator variants, we select Qwen2.5-0.5B-Instruct and Qwen3.5-0.8B from the Qwen series[6], and LLaMA-3.2-1B-Instruct and LLaMA-3.2-3B-Instruct from the LLaMA series[7]. These variants enable us to examine the effects of backbone family and model size on evaluator performance.

[2] https://www.12371.cn/special/20dbgxx/

[3] https://github.com/rangwang/CCAC2024-FS_Moderation

[4] https://huggingface.co/datasets/allenai/ValuePrism

[5] http://cvoices.net/

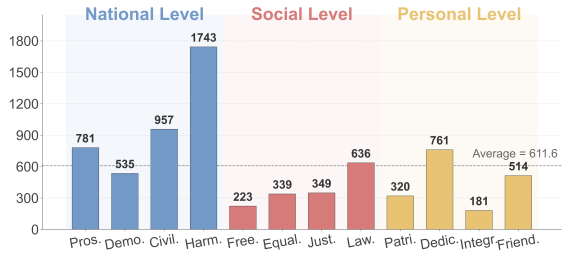
[6] https://qwen.ai/blog?id=qwen3

[7] https://www.llama.com/models/llama-3/

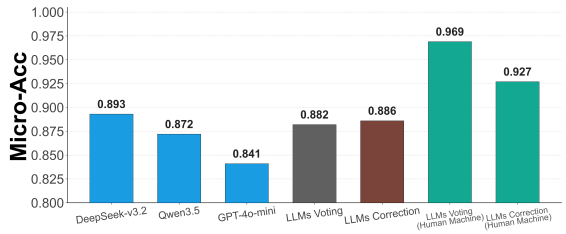
Evaluation Metrics. We use the metric of Micro-Accuracy (Micro-Acc) to evaluate the consistency between LLM-generated and human annotations, and use metrics of Accuracy and F1-Score to evaluate the classification performance of C-VoicesEvaluator across the 12 value dimensions.

Implementation. The C-Voices dataset is split into training, validation, and test sets at 8:1:1 ratio. During training, the number of epochs is set to 20 and the batch size is set to 8. Adam is adopted as the optimizer with a learning rate of $1e^{-5}$. Experiments are conducted on eight NVIDIA RTX A5000 GPUs. The hyperparameters of ASL loss function follow Ridnik et al. (Ridnik et al., 2021).

Annotation Cost. Human annotation costs 0.5 RMB per sample per annotator; annotating 9,000 candidate samples with three annotators costs 13,500 RMB. Manual preprocessing and filtering cost an additional 3,600 RMB. In comparison, LLM-based annotation is substantially cheaper, with per-sample costs of approximately 0.006, 0.005, and 0.007 RMB for DeepSeek-R1, GPT-4o-mini, and Qwen3.5, respectively. The two-round cross-correction annotation with DeepSeek-v3.2 and Qwen3.5 costs 96.2 RMB for 3,700 samples. Overall, constructing C-Voices costs approximately 17,200 RMB.



(a) Value distribution of human annotated C-Voices



(b) Micro-Acc performance of LLM-based annotations

Figure 4. Performance analysis of CSV annotations

5.2 Results of Value Annotation

5.2.1 Value Distribution of C-Voices.

Figure 4a illustrates the value distribution of the human-annotated C-Voices dataset. The 12 CSV dimensions are imbalanced: national-level values such as Prosperity and Democracy account for larger proportions, while social- and individual-level values such as Equality and Integrity are relatively sparse. This imbalance is partly due to the data source, People's Daily, whose reports naturally emphasize national-level narratives. To improve diversity, we adopt section-wise sampling across politics, economy, society, ecology, and other sections, followed by manual filtering. The resulting imbalance further motivates the use of asymmetric loss during evaluator training.

5.2.2 Performance of LLM-based Value Annotation.

We take human annotation as the benchmark and measure the consistency between LLM-generated and human annotations, as shown in Figure 4b.

Single-LLM annotation achieves an average Micro-Acc of 0.87, while majority voting and cross-correction improve it to 0.88 and 0.886, respectively. This shows that multi-model collaboration reduces prediction variance and improves annotation consistency. Entropy-based human-LLM collaboration further improves performance: the majority-voting-based human–LLM strategy reaches 0.96 Acc, an 11.5% improvement over the single-model baseline, confirming the effectiveness of human intervention for uncertain samples.

5.3 Performance of C-VoicesEvaluator

To further assess annotation quality, we train C-VoicesEvaluator on datasets produced by different annotation strategies. Table 3 reports results with Qwen2.5-0.5B-Instruct. The training and validation sets follow each annotation strategy, while all models share the same human annotated test set.

We have the following three observations. First, Correction LLMs, Triple LLMs, and DeepSeek-v3.2 perform comparably, while Qwen3.5 and GPT-4o-mini are weaker. All LLM-based strategies still lag behind the human benchmark, indicating that LLMs can support CSV annotation but are not yet fully reliable. Second, dimensions with clearer semantics, such as Rule of Law, Freedom, Harmony, and Patriotism, achieve higher F1 scores than more abstract or overlapping dimensions, including Equality and Justice. Third, F1-score is generally lower than Accuracy, mainly due to 1-0 imbalance in the multi-label setting, where models tend to predict dominant label 0.

Table 3. Experimental results of C-VoicesEvaluator based on Qwen2.5-0.5B-Instruct.

Level	Dimension	Single-LLM						Triple LLMs		Correction LLMs		Human Annotation	
		Qwen3.5		DeepSeek-v3.2		GPT-4o-mini							
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
National	Pros.	0.84	0.70	0.86	0.72	0.73	0.60	0.85	0.71	0.86	0.73	**0.89**	**0.74**
	Demo.	0.89	0.47	0.88	0.40	0.89	0.52	0.88	0.46	0.88	0.40	**0.91**	**0.67**
	Civil.	0.86	0.71	0.85	0.70	0.85	0.68	0.86	0.70	0.85	0.70	**0.87**	**0.74**
	Harm.	0.74	0.69	0.75	0.69	0.74	0.76	0.74	0.72	0.75	0.73	**0.80**	**0.79**
	Avg.	0.83	0.64	0.84	0.63	0.80	0.64	0.83	0.65	0.84	0.64	**0.87**	**0.74**
Social	Free.	0.94	0.63	0.97	0.73	0.97	0.78	0.97	0.72	**0.98**	0.75	0.98	**0.79**
	Equal.	0.91	0.50	**0.93**	0.50	0.91	0.53	0.91	0.48	**0.93**	0.45	0.93	**0.55**
	Just.	0.91	0.59	0.93	0.54	0.89	0.57	0.92	0.59	0.92	**0.61**	**0.94**	**0.61**
	Law.	0.92	0.79	0.93	**0.81**	0.91	0.79	**0.93**	0.80	**0.93**	0.80	**0.93**	0.80
	Avg.	0.92	0.63	0.94	0.65	0.92	0.67	0.93	0.65	0.94	0.65	**0.95**	**0.69**
Personal	Patri.	0.91	0.61	0.95	0.68	0.89	0.57	0.93	0.65	0.93	0.68	**0.96**	**0.76**
	Dedic.	0.82	0.65	0.87	0.66	0.67	0.54	0.84	0.68	0.86	0.69	**0.88**	**0.70**
	Integr.	**0.97**	0.74	**0.97**	**0.75**	**0.97**	**0.75**	**0.97**	0.70	**0.97**	0.72	**0.97**	0.70
	Friend.	0.88	0.57	0.89	0.47	0.85	0.51	0.88	0.55	0.90	0.56	**0.93**	**0.74**
	Avg.	0.90	0.64	0.92	0.64	0.85	0.59	0.91	0.65	0.92	0.66	**0.94**	**0.73**

Table 4. Impact of model size on C-VoicesEvaluator

Model	Metric	Pros.	Demo.	Civil.	Harm.	Free.	Equal.	Just.	Law	Patri.	Dedic.	Integr.	Friend.	Avg.
LLaMA-3.2-1B-Instruct	ACC.	0.88	0.88	**0.87**	**0.77**	0.98	0.92	**0.93**	0.91	0.95	0.85	**0.97**	0.93	0.90
	F1	0.74	**0.53**	0.70	**0.77**	0.61	0.55	**0.73**	0.79	0.61	0.65	**0.73**	0.74	0.68
LLaMA-3.2-3B-Instruct	ACC.	**0.89**	**0.89**	**0.87**	**0.77**	**0.99**	**0.94**	0.92	**0.94**	**0.96**	**0.87**	0.96	**0.94**	**0.91**
	F1	**0.78**	0.51	**0.71**	**0.77**	**0.76**	**0.63**	0.67	**0.85**	**0.67**	**0.66**	0.67	**0.76**	**0.70**

5.4 Performance of Value Alignment Evaluation

This section evaluates eight mainstream LLMs using C-VoicesEvaluator trained on Qwen2.5-0.5B-Instruct. Figure 5 shows their CSV distributions.

The models differ substantially across the 12 CSV dimensions. In Figure 5, the gray region at 0 denotes the C-Voices human benchmark; positive values indicate stronger expression than the benchmark, and negative values indicate weaker expression. Kimi-K2.6, Gemini-2.5-Pro, and DeepSeek-V4-Pro show smaller deviations, suggesting closer alignment with C-Voices. In contrast, GPT-4.1 and LLaMA-4-Scout show stronger tendencies in Freedom, Equality, and Integrity, but weaker tendencies in Prosperity and Dedication. Most models also show high deviations on Freedom, partly because Freedom-related samples are sparse in C-Voices, amplifying relative deviation values.

5.5 Ablation Study

This section analyzes the effects of model size and annotation data scale on C-VoicesEvaluator.

Impact of Model Size. Table 4 compares evaluator performance with different backbone sizes. LLaMA-3.2-3B-Instruct slightly outperforms LLaMA-3.2-1B-Instruct, improving average Accuracy from 0.90 to 0.91 and average F1 from 0.68 to 0.70. It also achieves higher F1 scores on most dimensions, with the largest gain on Freedom, where F1 increases from 0.61 to 0.76. These results show that larger backbones bring moderate improvements, but the gains tend to saturate as model size increases.

Impact of Annotated Data Scale. Figure 6 shows the effect of annotated data scale on evaluator performance, where C-VoicesEvaluator is trained with different subsets of C-Voices. As the training size increases from 1,000 to 3,500 samples, the average

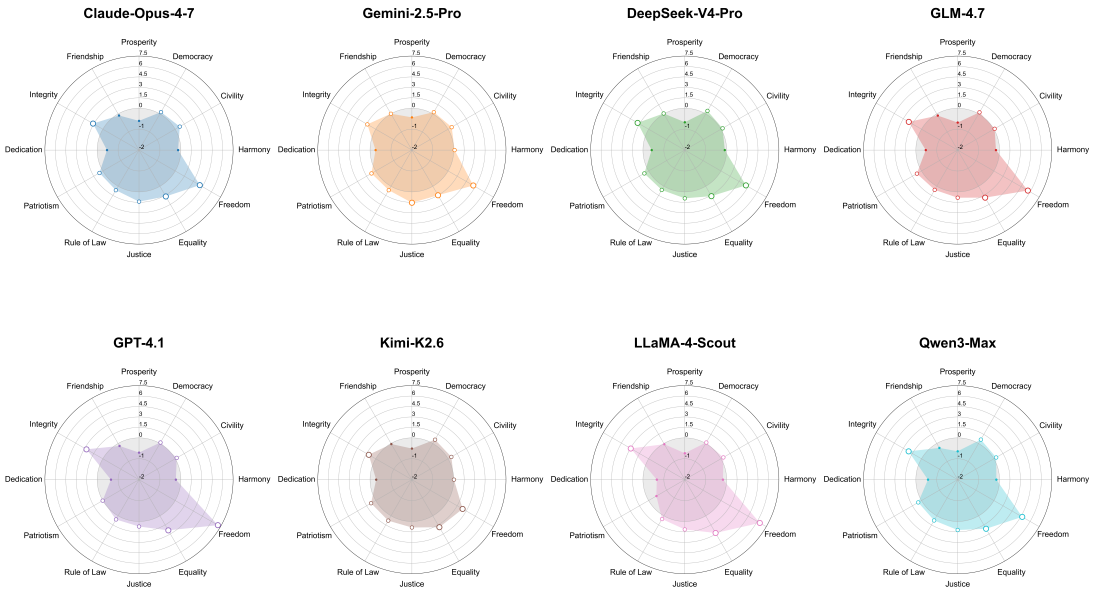


Figure 5. Radar charts of CSV assessment for different LLMs

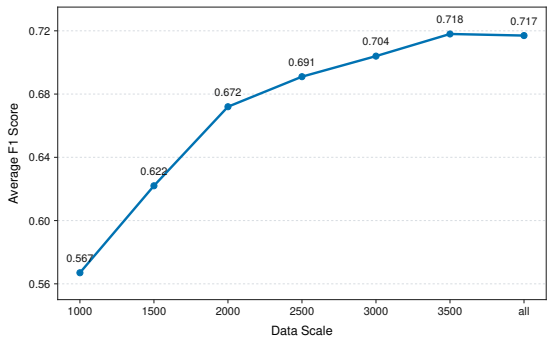


Figure 6. Impact of annotation data scale on C-VoicesEvaluator

F1 score steadily improves from 0.567 to 0.718. Adding more data beyond 3,500 samples brings only marginal gains, with performance using all 3,700 samples remaining nearly unchanged. These results suggest that the 3,700 annotated samples are sufficient for training an effective evaluator.

6 Conclusion

This paper proposes **C-VoicesCompass**, a CSV-based value alignment evaluation framework for LLMs. We construct **C-Voices**, the first annotated dataset for CSV, and train **C-VoicesEvaluator**, a lightweight evaluator with asymmetric loss. We further examine LLM-based value annotation and show that multi-round cross-correction achieves performance close to human annotation. Experiments validate the effectiveness of C-Voices and C-VoicesEvaluator, and reveal clear alignment differences among mainstream LLMs under the Chinese Social Values framework.

Limitations

C-Voices currently focuses on Chinese news reports, which provide value-rich and high-quality content but do not cover all domains and communication scenarios. Future work could extend the dataset to conversational, social-media, and multimodal contexts. Second, the current framework represents the 12 CSV dimensions using binary labels. Although this design enables consistent and scalable evaluation, graded and context-sensitive annotations may better capture value intensity and conflicts. Finally, C-VoicesCompass is intended to complement, rather than replace, general-purpose value and safety evaluations.

References

Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

Yuntao Bai, Andy Jones, Kamal Ndousse, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081.

Georgios Chochlakis, Peter Wu, Arjun Bedi, Marcus Ma, Kristina Lerman, and Shrikanth Narayanan. 2025. Humans hallucinate too: Language models identify and correct subjective annotation errors with label-in-a-haystack prompts. *Preprint*, arXiv:2505.17222.

Arduin Findeis, Floris Weers, Guoli Yin, Ke Ye, Ruoming Pang, and Tom Gunter. 2025. Can external validation tools improve annotation quality for llm-as-a-judge? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15997–16020.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Fabrizio Gilardi, Meysam Alizadeh, and Ma"el Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Michael Gow. 2017. The core socialist values of the Chinese dream: Towards a Chinese integral state. *Critical Asian Studies*, 49(1):92–116.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey: Round seven—country-pooled datafile version 5.0. JD Systems Institute and WVSA Secretariat.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190. Association for Computational Linguistics.

Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2024. Flames: Benchmarking value alignment of LLMs in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4551–4591. Association for Computational Linguistics.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah G"um"uş, and Michael Granitzer. 2024. LLMs in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, pages 397–412. Springer Nature Switzerland.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213. Association for Computational Linguistics.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on Hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503. Association for Computational Linguistics.

Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Miserlis Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2026. Can reasoning help large language models capture human annotator disagreement? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–54.

Abiodun Finbarrs Oketunji, Muhammad Anas, and Deepthi Saina. 2023. Large language model (LLM) bias index—LLMBI. *Preprint*, arXiv:2312.14769.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Anh C. Pham, Mihir Thalanki, Michael Sun, Aditya Chaloo, Ankita Gupta, Tian Xia, Aditya Mate, Ehi Nosakhare, and Soundararajan Srinivasan. 2025. How to fine-tune safely on a budget: Model adaptation using minimal resources. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1970–1981. Association for Computational Linguistics.

Xiangyu Qi, Yi Zeng, Tinghao Xie, and 1 others. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *Preprint*, arXiv:2310.03693.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119. Association for Computational Linguistics.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Computing Surveys*, 54(9):1–40.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating llm-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Academic Press.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Li Song. 2021. Construction of socialist core values from Chinese traditional culture. *International Journal of Frontiers in Sociology*, 3(12):69–76.

UNESCO. 2021. Recommendation on the ethics of artificial intelligence. Technical report, United Nations Educational, Scientific and Cultural Organization, Paris, France.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025a. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6859–6877. Association for Computational Linguistics.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025b. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8762–8785. Association for Computational Linguistics.

Yuchuan Zhang. 2013. Reflections on the condensation of Chinese social values. *Journal of Chongqing Technology and Business University (Social Science Edition)*, 30(3):52–58.