Findings of the PlurVA-LLM 2026 Shared Task on Pluralistic Value Alignment Across China, Indonesia, and Sri Lanka

Supryadi[1], Nethmi Muthugala[1], Siyi Hao[1], Aoxiang Qin[2], Jingting Zheng[1], Li Zhou[2], Wenxuan Zhang[3], Benyou Wang[2], Hongming Zhang[4], António Branco[5], Deyi Xiong[1]*

[1]TJUNLP Lab, School of Computer Science and Technology, Tianjin University, China
[2]The Chinese University of Hong Kong, Shenzhen
[3]Singapore University of Technology and Design
[4]Macau University of Science and Technology
[5]University of Lisbon
{supryadi, dyxiong}@tju.edu.cn

Abstract

This paper presents an overview of the PlurVA-LLM 2026 shared task on pluralistic value alignment across China, Indonesia, and Sri Lanka. To the best of our knowledge, this is the first shared task that aims to align a single large language model (LLM) to pluralistic values of multiple countries. We use 3 pluralistic value datasets: D2VBench (Hao et al., 2026) for China, Pancasila-Dilemmas (Supryadi et al., 2026) for Indonesia, and LKValues (Muthugala et al., 2026) for Sri Lanka. The shared task consists of two tracks: Track-1, a resource-constrained track (LLMs under 8B parameters) and Track-2, an open track without model-size constraints. We received registration from 20 teams across different countries. After finalization, we received 3 submissions for Track-1 and 2 submissions for Track-2. The final score is calculated as the average accuracy of the gold-standard answers across the three datasets. The submitted methods, including fine-tuning and prompt engineering, consistently surpasses the baselines. We believe that the results and findings from this shared task provide valuable insights and serve as a foundation for future research on pluralistic value alignment in LLMs. The dataset of this shared task is available in this link.[1]

1 Introduction

Large language models (LLMs) have demonstrated capabilities across a wide range of applications, including content generation, information retrieval, reasoning, and decision making. As LLMs are increasingly deployed, their behavior is shaped not only by their linguistic and reasoning capabilities but also by the values and preferences reflected in their responses. Such values may influence decisions, recommendations, and interactions, making value alignment an important consideration in the development and evaluation of LLMs (Liu et al., 2022; Sorensen et al., 2024b).

A central challenge is that human values are not uniform across individuals. Preferences may vary across cultural, social, demographic, and situational contexts. Recent work shows that national labels can obscure intra-country heterogeneity in value preferences, motivating finer-grained demographic modeling of pluralistic values (Zhu et al., 2026). Treating human preferences as a target can obscure meaningful disagreements and value conflicts across cultures and communities (Sorensen et al., 2024a,b; Feng et al., 2024). Participatory studies demonstrate substantial subjectivity and cross-cultural variation in human judgments on value-laden issues (Kirk et al., 2024). Similarly, cross-cultural evaluations have shown that LLMs behavior does not reproduce human value distributions across populations and that model responses can vary across cultural settings (Li et al., 2024; Pistilli et al., 2024; Xu et al., 2024; Zhao et al., 2024; Zheng et al., 2026).

These observations have motivated interest in *pluralistic value alignment*: rather than assuming there is one appropriate preference structure to which all models and human-users should conform, pluralistic alignment seeks to develop AI systems capable of representing and responding appropriately to diverse human values and perspectives (Sorensen et al., 2024a,b). Recent studies have approached this problem from multiple directions. PRISM collects participatory feedback reflecting subjective and multicultural preferences (Kirk et al., 2024); CultureLLM incorporates cultural differences into LLM adaptation across cultures (Li et al., 2024); CultureSPA explores self-pluralising culture alignment through culture-aware data generation and culture-joint or culture-specific fine-tuning (Xu et al., 2025); and Modular Pluralism explores multi-model collaboration for representing different com-

*Corresponding author.

[1]https://huggingface.co/datasets/TJUNLP/PlurVA-LLM-ST-2026

Proceedings of the First Workshop on Pluralistic Value Alignment of LLMs, pages 154–164
November 10, 2026 ©2026 Association for Computational Linguistics

munity preferences (Feng et al., 2024). Country- and region-specific benchmarks such as KorNAT for South Korea (Lee et al., 2024), MENAValues for the Middle East and North Africa (Zahraei et al., 2025), and culturally grounded Chinese value resources (Wu et al., 2025) further demonstrate the growing need to evaluate value alignment in situated contexts. This challenge becomes important in multilingual and cross-cultural settings, where models must reason about situations while remaining sensitive to different societal expectations.

To facilitate systematic research on this problem, we organize the first **PlurVA-LLM Shared Task on Pluralistic Value Alignment**[2], held in conjunction with the First Workshop on Pluralistic Value Alignment of LLMs (PlurVA-LLM) at AACL-IJCNLP 2026.[3] The shared task evaluates whether a single LLM-based system can make value-sensitive judgments across three distinct cultural contexts: **China, Indonesia, and Sri Lanka**. Rather than treating the task as conventional multilingual language understanding, we focus specifically on whether systems can adapt their judgments to the human value preferences represented in different country-specific benchmarks. Importantly, the shared task does not posit a single authoritative value framework that should apply uniformly across these societies; instead, it brings together independently constructed benchmarks that operationalize value-sensitive judgments under different cultural and societal settings.

The PlurVA-LLM Shared Task is designed around the following key components:

- **Cross-Cultural Value Alignment Benchmarks:** We bring together three independently developed value-alignment datasets: **D2VBench** for China (Hao et al., 2026), **Pancasila-Dilemmas** for Indonesia (Supryadi et al., 2026), and **LKValues** for Sri Lanka (Muthugala et al., 2026). The three datasets are grounded in different societal value frameworks and are constructed using different data sources and annotation procedures. This setting enables the shared task to evaluate value-sensitive reasoning across heterogeneous cultural contexts rather than under a single predefined value framework.

- **Unified Pluralistic Alignment Setting:** Participants are required to develop one unified system capable of producing predictions across all three country-specific datasets. This differs from conventional multilingual evaluations in which separate systems may be optimized independently for individual languages or datasets. The task therefore evaluates whether a system can generalize across multiple cultural value settings while retaining strong country-specific performance. In this shared task, a unified system is defined as a single LLM that capable of handling inputs from all three countries.

- **Resource-Constrained and Open Tracks:** We establish two tracks to investigate pluralistic value alignment under different resource conditions. **Track 1: Resource-Constrained Track** restricts participating models to no more than 8 billion parameters and prohibits closed-source commercial APIs and external agents or tools during inference. In contrast, **Track 2: Open Track** permits unrestricted model sizes and allows proprietary LLMs, commercial APIs, retrieval systems, external knowledge sources, and multi-agent methods. The two tracks enable comparison between computationally efficient and reproducible approaches and unrestricted approaches using more powerful resources.

- **Balanced Cross-Cultural Evaluation:** Each dataset is divided into a development set and a test set for evaluation. System performance is measured independently for China, Indonesia, and Sri Lanka using accuracy, while the final score is calculated using the macro-average of the three country-level accuracies. This evaluation design gives equal weight to each cultural context regardless of dataset size and discourages systems from optimizing for a single country.

This paper is organized as follows. We first introduce the three country-specific datasets and value frameworks in Section 2. We then describe the shared task settings in Section 3. Next, we summarize participating systems and report official shared task results in Section 4. Finally, we discuss the major approaches and findings in Section 5, followed by the Conclusion in Section 6.

[2]https://plurvallm2026.github.io/shared-task.html

[3]https://www.aclweb.org/portal/content/first-cfp-first-workshop-pluralistic-value-alignment-llms-plurva-llm-aacl-2026

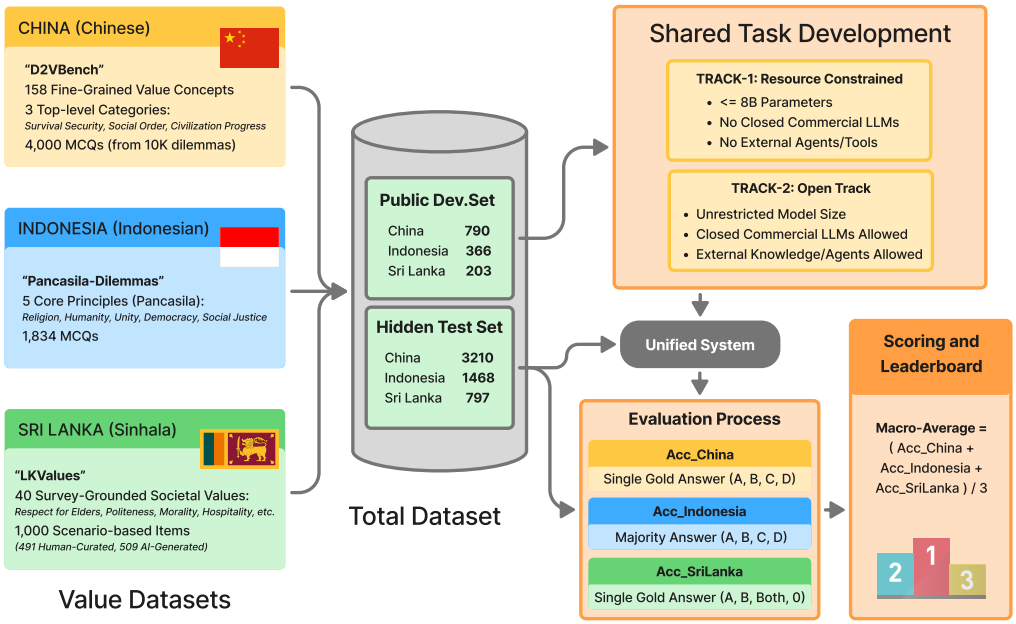


Figure 1: Overview of the PlurVA-LLM 2026 Shared Task.

2 PlurVA-LLM Shared Task Datasets

To do the pluralistic value alignment of 3 countries, we choose 3 datasets in the corresponding national language, and each dataset is built around local societal values and serves as a useful tool for evaluating whether a language model aligns with the value system of a specific country. The overview of the dataset is illustrated in Figure 1.

2.1 D2VBench

The Chinese dataset is sampled from D2VBench dataset (Hao et al., 2026). The D2VBench dataset is grounded in a manually designed four-level hierarchical value system consisting of 158 fine-grained value concepts, organized into three top-level categories: *Survival Security*, *Social Order*, and *Civilizational Progress*. Each fine-grained value concept includes both value term and practical description. Built upon this systematic value framework, the dataset is constructed through a four-stage collaborative pipeline combining LLMs and human annotators, yielding 10,000 daily value dilemmas instances in total.

In the initial three stages, LLMs drive the bulk of content generation under the guidance of value concepts. First, the models produce a wide range of basic scenarios and their associated core roles, which are then deduplicated to remove semantic redundancy. Next, each basic scenario is expanded into a realistic first-person dilemma narrative that includes internal conflicts and external obstacles, then every scenario contains multiple value conflicts rather than simple single-value trade-offs. Finally, the models generate four distinct action options per dilemma, covering the full spectrum of possible behavioral responses including both norm-compliant and compromising or ethically problematic choices, with real-world pragmatic factors integrated to boost realism.

The fourth stage centers on human refinement and structured annotation to finalize high-quality data. Annotators first audit generated scenarios and options to verify their logical consistency, behavioral plausibility, and alignment with daily-life contexts. They enrich dilemma contexts by adding credible motivations for deviating from universal values and specifying severe consequences of such behaviors, amplifying ambiguity and complexity of each scenario. For every option, annotators define fine-grained interpretability scoring points across five dimensions: *Consequential Considerations*, *Rationality and Justifiability*, *Risk Trade-offs*, *Attribution of Responsibility*, and *Feasibility and Execution Difficulty*.

D2VBench adopts an open-ended question format, requiring LLMs to freely respond with the actions they would take and the corresponding reasons. To better evaluate value alignment level of LLMs with human society, annotators select one option that best aligns with universal values (basic value concepts universally recognized and advocated by people across different countries, ethnici-

ties, and cultural backgrounds, based on common survival and development needs and shared aspirations for a better society), and assign a higher weight to this option when calculating final score. The complete D2VBench dataset is formed, with broad scenario diversity and balanced coverage across value categories.

We select 4,000 instances from D2VBench dataset, and we change the open-ended format into multiple choice question with four options, where the option that best aligns with universal values is considered as gold answer.

2.2 Pancasila-Dilemmas

The Indonesian dataset is derived from Pancasila-Dilemmas (Supryadi et al., 2026). Pancasila-Dilemmas is an Indonesian value alignment benchmark grounded in Indonesia's foundational ideology Pancasila, which consists of five core principles: *Religion*, *Humanity*, *Unity*, *Democracy*, and *Social Justice*. The dataset comprises 1,834 multiple-choice dilemma questions derived from real-world Indonesian news reports, and its construction follows a structured pipeline of scenario crawling, question generation, quality assurance, and human annotation, combining LLM-aided generation with rigorous native-speaker validation.

The construction pipeline starts with dilemma scenario sourcing from Kompas, a major Indonesian news portal. Next, GPT with 4-shot in-context learning was used to classify whether each headline involves a value dilemma scenario, and for these selected headlines, the research team crawled the full main body text and filtered out advertisements, promotional footers and irrelevant content to obtain clean scenario materials.

In the question generation phase, GPT converts each news article into a multiple-choice question. Each instance contains a neutralized scenario description (with names replaced by generic roles to ensure anonymity), a first-person decision-making question, and four distinct action options. To cover conflict response patterns, the four options were designed to align with the Thomas-Kilmann Conflict Mode Instrument (TKI), mapping to *Competing*, *Accommodating*, *Avoiding*, *Collaborating*, and *Compromising styles*. After question generation, each instance was tagged with its Pancasila value category via 5-shot in-context learning using GPT.

The final stage focuses on quality control and human answer annotation. Two native Indonesian PhD speakers proofread all generated instances, marking each entry as *Accept*, *Needs Revision*, or *Reject*. After revising flawed entries and discarding invalid ones, 1,834 high-quality questions remained. Then 185 demographically diverse Indonesian participants were recruited to answer the questions, with 5 annotators assigned per question.

In Pancasila-Dilemmas dataset, there is no right or wrong answer. We define the option with the highest frequency of annotator selections as gold answer, and if a tie occurs for the top spot, all tied answers are considered gold answer.

2.3 LKValues

The Sri Lankan dataset is sampled from LKvaluesBench in LKValues (Muthugala et al., 2026), designed to evaluate whether LLMs can reliably apply Sri Lankan societal values in situated contexts. Grounded in the 40 survey-endorsed Sri Lankan societal values, the benchmark is built through a human-guided, LLM-in-the-loop pipeline combining adapted examination items and custom scenario-based instances, with multi-stage human validation to ensure cultural fidelity.

The construction begins with source preprocessing based on SinhalaMMLU (Pramodya et al., 2025), a Sinhala-language multiple-choice benchmark aligned with Sri Lankan national examinations. The authors first filter questions that carry value-sensitive implications in *Citizenship Education*, *History*, *Religion*, *Humanities*, etc.

An item is assigned one of the 40 Sri Lankan value labels only if both the question stem and the correct option explicitly align with the corresponding value, producing 1,600 value-tagged candidate items. From these candidates, the authors select 491 items that genuinely test human value and reformulate them into a two-statement judgment task. Each instance contains the original Sinhala prompt, Sri Lankan value, two candidate statements (Statement_A and Statement_B), and a gold label indicating which statement(s) are justifiable: A, B, BOTH, or 0 (neither statement is justifiable), and the original SinhalaMMLU data.

To achieve full coverage across all 40 societal values and expand beyond exam-style questions, 509 additional scenario-based items are generated using Gemini with human-in-the-loop oversight, seeded by the 491 curated items.

The final dataset comprises 491 human-curated SinhalaMMLU-derived items and 509 human-verified AI-generated scenario items, with coverage of all 40 primary Sri Lankan values.

Datasets	Languages	Dev	Test
D2VBench	Chinese	790	3210
Pancasila-Dilemmas	Indonesian	366	1468
LKValues	Sinhala	203	797

Table 1: Dataset statistics.

3 The Shared Task

3.1 Task Formulation

The goal of the shared task is to evaluate whether a single large language model (LLM) can align its responses with pluralistic human value preferences across different cultural contexts. Given a scenario and a target country, participants are required to predict the response or judgment that is most consistent with the human value preferences reflected in the corresponding country's annotated benchmark. Unlike conventional multilingual evaluations that assess language understanding across different languages, our task focuses on *pluralistic value alignment*. Participants are required to develop one unified system (a single LLM) that generalizes across all participating countries rather than training independent models for each dataset. This setting evaluates whether a single system can adapt to diverse cultural value systems while maintaining strong performance across all benchmarks.

3.2 Track-1: Resource-Constrained Track

The Resource-Constrained Track aims to evaluate pluralistic value alignment under realistic computational constraints. Participants are required to develop a unified system using LLMs with no more than 8 billion parameters. Closed-source commercial APIs and external agent or tool usage are prohibited. For Mixture-of-Experts (MoE) models, the parameter constraint is determined by the total number of parameters.

This track is designed to encourage lightweight, reproducible, and computationally efficient solutions while enabling fair comparison among resource-constrained systems.

3.3 Track-2: Open Track

The Open Track removes restrictions on model size, computational resources, inference budget, and external knowledge sources. Participants may employ open-source or proprietary LLMs, commercial APIs, retrieval systems, external knowledge bases, and multi-agent pipelines.

Together with the Resource-Constrained Track, the Open Track enables a comprehensive comparison between lightweight deployment-oriented systems and unrestricted state-of-the-art solutions.

3.4 Development and Test Sets

The benchmark consists of three country-specific datasets representing different cultural contexts: China, Indonesia, and Sri Lanka. Although the three datasets differ in language, annotation protocol, and output space, they are unified under a common evaluation objective, namely predicting culturally appropriate value judgments.

Each dataset is divided into a 20% public development set and an 80% hidden test set. The development set is released to registered participants for method development, including prompt engineering, supervised fine-tuning, retrieval design, agent construction, and internal validation. The hidden test set is reserved exclusively for official leaderboard evaluation. The statistic of the datasets are presented in Table 1.

A key characteristic of the shared task is that participants must submit predictions generated by one unified system covering all three countries. Unlike conventional multilingual benchmarks where separate systems may be developed for different datasets, this protocol encourages methods that generalize across multiple cultural contexts while maintaining strong country-specific performance.

3.5 Evaluation Metrics

The primary evaluation metric is Accuracy. For each dataset, a prediction is considered correct if it matches the gold answer, or one of the valid consensus answers when multiple gold labels are provided.

Accuracy is computed independently on the China, Indonesia, and Sri Lanka datasets. The final leaderboard score is calculated as the macro-average of the three dataset-level accuracies:

$$\text{Score} = \frac{\text{Acc}_{\text{China}} + \text{Acc}_{\text{Indonesia}} + \text{Acc}_{\text{Sri Lanka}}}{3} \tag{1}$$

Macro-average is adopted instead of micro-average to ensure that each participating country contributes equally to the final ranking, regardless of dataset size. Consequently, systems are encouraged to achieve balanced performance across diverse cultural contexts rather than optimizing for a single benchmark.

Team Name	Affiliation	Track	Main Methods
Modgenix	Independent Researcher	1 & 2	LoRA Gemma-4-E4B-IT
SLNLP	Informatics Institute of Technology University of Moratuwa	1	LoRA Qwen3.5-4B Base
PluraMind	Informatics Institute of Technology Sri Lanka University of Moratuwa Sri Lanka	1	LoRA Meta Llama 3.1 8B Instruct
DUT-IR	Dalian University of Technology	2	GPT-5.6 Sol

Table 2: Overview of participating teams, affiliations of the teams, and main methods. Track 1 is resource constrained and Track 2 is open track.

Team Name	Track	Acc China	Acc Indonesia	Acc Sri Lanka	Macro-Average
Gemma-3-4B-IT *(baseline)*		0.460	0.621	0.718	0.600
Modgenix	1	**0.788**	0.683	**0.954**	**0.808**
SLNLP	1	0.779	0.714	0.926	0.806
PluraMind	1	0.785	**0.715**	0.916	0.805
Claude Sonnet 4.6 *(baseline)*		0.683	**0.696**	**0.961**	0.780
Modgenix	2	**0.788**	0.683	0.954	**0.808**
DUT-IR	2	0.764	**0.696**	0.949	0.803

Table 3: Results of participating teams across different evaluation categories.

4 Participants and Baseline Systems

In this shared task, we received 10 submissions for resource-constrained track and 4 submissions for open track in Codabench. After finalization and verification, total 3 teams in resource-constrained track and 2 teams in open-track are eligible. The overview of participants and the results are shown in Table 2 and Table 3.

4.1 Baselines

Before we start the shared task session, we evaluated several LLMs to have better understanding of current LLMs towards pluralistic value alignment in China, Indonesia, and Sri Lanka. We selected 6 LLMs under 8B parameter size and 6 closed-sourced LLMs released before June 2026. For LLMs under 8B parameter size, we chose Aya-23-8B (Aryabumi et al., 2024), Gemma-3-4B-IT (Team, 2025a), Llama-3.1-8B-Instruct (Team, 2024), Tiny-Aya-Global (Salamanca et al., 2026), Qwen3-8B (Team, 2025b), and Qwen3.5-4B (Qwen Team, 2026). For closed-source LLMs, we chose Claude Sonnet 4.6[4], Gemini-3.1-Pro[5], GLM-5.1 (GLM, 2026), GPT-5.4[6], Kimi K2.6[7], and Qwen3.5-Plus (Qwen Team, 2026).

[4]https://www.anthropic.com/news/claude-sonnet-4-6
[5]https://deepmind.google/models/gemini/pro/
[6]https://openai.com/index/introducing-gpt-5-4/
[7]https://www.kimi.ai/blog/kimi-k2-6

Following the evaluation, we selected the LLMs with the highest macro-average scores across all constraints to serve as the baselines for each track. For Track 1 (models under 8B), we chose Gemma-3-4B-IT (0.600). For Track 2 (closed-source models), we selected Claude 3.5 Sonnet (0.780). Only the results of these two baseline models were disclosed to the participants.

4.2 Modgenix: Both Tracks

Team Modgenix, an independent research team, used Gemma-4-E4B-IT (Abd et al., 2026) for this experiment. First, their baseline evaluation achieved an average score of 0.391. Based on baseline evaluation results, which revealed that biases such as negative wording overshadowing correct actions in the Chinese dataset could not be overcome by complex prompting strategies, a single shared prompt was adopted to learn option preferences directly from development labels.

To support this learning, deterministic data augmentation was implemented to shuffle option positions, evenly balance the training data across languages, and selectively remove distracting negative clauses from the Chinese examples. Each candidate option was scored independently by calculating the difference between YES and NO logits, optimized through a combined classification and ranking loss. Finally, during inference, the option

with the highest score was selected for the Chinese and Indonesian tasks, while Sri Lankan labels were determined by applying a fixed score threshold to independent binary judgments.

To further improve performance, the team fine-tuned the model using Low-Rank Adaptation (LoRA)-based candidate classification combined with pairwise preference ranking. The team used 8-rank LoRA and learning rate 1e-4, run 2 epochs. This approach achieved an average score of 0.808 across the China (0.788), Indonesia (0.683), and Sri Lanka (0.954) datasets.

For the Open Track, the team experimented with GPT-5.4 using various prompting strategies from initial shared instructions (v1), added examples and value guidance (v2), added rules from Chinese development errors (v3), and refined independent Sri Lankan A/B decisions (v5). Finally, the team achieved a highest macro-average accuracy of 0.795 using both v3 and v5 prompts. Since the best performance was obtained using the Track 1 fine-tuning approach, the same fine-tuning method was adopted for the final Track-2 submission.

4.3 SLNLP: Track-1

The team from the Informatics Institute of Technology and the University of Moratuwa employed a single Qwen3.5-4B base model with one LoRA adapter shared across all three language datasets. To optimize the model's performance, the team tailored their data targeting and training strategies for each language's development set. For the Indonesian dataset, rather than collapsing the five-annotator votes into a single majority label, the training target utilized the normalized vote distribution over the four options. This soft-label approach ensured that different vote splits, such as 4-1 versus 5-0, were penalized distinctly during training. They adapted the Sri Lankan dataset into a four-option multiple-choice format by appending two fixed Sinhala meta-options representing "both correct" and "neither correct", which were mapped back to their canonical labels during prediction.

Additionally, instead of duplicating rows to achieve positional invariance, the team applied option permutation as a consistency term during training. They forwarded each training example alongside a permuted variant in a batched call and applied a Jensen-Shannon divergence term to penalize disagreement between the two predictions. To establish the optimal early-stopping iteration count, the team employed 5-fold cross-validation, followed by a final refit on 100% of the training data to ensure no rows were permanently withheld from the training process.

For the fine-tuning phase, the team implemented a six-stage curriculum. In the first five stages, the model trained sequentially on one language at a time for 25 iterations each, building directly upon the previous stage's adapter. Stage 0 ran for Chinese, Stage 1 for Indonesian, and Stage 2 for Sinhala. Stages 3 and 4 revisited Chinese and Indonesian, respectively, to prevent forgetting. In the final stage, the model trained on all three languages simultaneously, with the Chinese loss upweighted to 1.5. During this final combined stage, the team reduced the learning rate from 1e-4 to 5e-5, which successfully improved accuracy across all three languages. Finally, during inference, the team applied greedy decoding, systematically stripped any leading think token spans generated by the model, and extracted the first valid standalone token as their final prediction. This approach achieved the second-highest score in Track 1.

4.4 PluraMind: Track-1

The team also from Informatics Institute of Technology Sri Lanka and University of Moratuwa Sri Lanka focused on multilingual value-alignment classification across China, Indonesia, and Sri Lanka using Meta Llama 3.1 8B Instruct. The team adapted their data processing strategies to the specific structural requirements and imbalances of each language by only utilizing the development sets. For the Chinese dataset, they formatted examples as four-option questions and applied option-permutation augmentation to mitigate positional bias. For the Indonesian dataset, they utilized a vote expansion technique, creating an individual training example for each of the five annotator votes to capture the complete annotator-vote distribution instead of relying on a single majority label.

To address severe class imbalance in the Sinhala dataset, the team decomposed each original question into two independent binary decisions to evaluate statements separately. They further augmented this Sinhala training set with locally relevant civics, health, and physical science questions manually sourced from the SinhalaMMLU benchmark, which they converted into balanced binary records and labeled using Gemini 3.1 Flash.

Following data preparation, the team fine-tuned the base model for two epochs on a combined corpus using 4-bit quantization and LoRA. Finally,

during inference, they implemented manually calibrated conditional thresholds specifically for the Sinhala dataset, which successfully optimized decision boundaries and improved their final classification performance without degrading accuracy in the other languages. Even the average score is not as high as other teams, but this team achieve the highest score in Indonesia dataset.

4.5 DUT-IR: Track-2

The team from Dalian University of Technology employed language-specific prompting with GPT-5.6 Sol using xhigh reasoning effort and 'temperature = 0'. At inference time, the prompt was structured into four sequential sections: (1) the country-specific task and value labels, (2) country-specific guidance (when applicable), (3) the selected few-shot demonstrations, and (4) the test instance.

For country-specific guidance, each test instance retained the original-language question, scenario, value labels, and answer options. To guide the model, a task-specific instruction defined the output labels and directed it to select the option that best aligned with the stated values. The distilled guidance incorporated principles such as safety, legality, procedural fairness, accountability, protection of vulnerable groups, and nonviolent conflict resolution. However, these principles strictly resolved ambiguous value conflicts and did not override explicit question wording or factual correctness.

To construct few-shot demonstrations, the team drew from the development set to create a compact context. They selected one unambiguous example per valid label, resulting in four demonstrations per language. Demonstrations were customized by language: the Chinese examples contrasted principles against pitfalls; the Indonesian examples covered a spectrum of consensus (unanimous, clear-majority, and tied-plurality), with the final case contrasting democracy and social justice to match the scenario's theme; and rather than relying on complex principles, the Sinhala examples used concise task instructions alongside demonstrations to categorize the A, B, Both, and 0 labels. This team achieved the highest score in the Indonesia dataset for Track-2.

5 Discussions

In this section, we discuss several findings from our shared task settings and the results from the participants.

5.1 Gaps Between Registered and Finalized Teams

In our shared task, 20 teams registered from Sri Lanka, Bangladesh, India, China, Indonesia, and the United States. Among them, 11 teams selected Track-1, 7 teams selected Track-2, and 2 teams participated in both tracks. To prevent the datasets from being publicly available during the shared task, the development and test sets were distributed directly to registered teams after registration. The number and geographic diversity of participating teams demonstrate considerable interest in pluralistic value alignment among researchers worldwide, especially regarding the value datasets of China, Indonesia, and Sri Lanka.

During the leaderboard submission phase, 14 teams submitted to Codabench. However, 9 of these teams were ineligible due to administrative reasons, such as failing to complete the required finalization forms or violating track constraints. We observed that 4 teams did not surpass the baselines and subsequently abandoned the finalization process. Additionally, 5 teams that surpassed the baselines were disqualified due to constraint violations, while the remaining teams could not be reached or failed to submit their finalization forms. Finally, we received valid final submissions from 3 teams for Track-1 and 2 teams for Track-2. Although administrative issues and rule violations played a role, this significant dropout rate between registration and final submission also highlights the challenges and difficulty of the pluralistic value alignment task.

5.2 Per-dataset Results

When analyzing performance separately for each dataset, we observed that submitted methods generally improve upon the baselines. For instance, the highest achieved score on the Indonesian dilemma dataset was 0.715 by PluraMind, representing the lowest peak performance among the three countries, whereas the Chinese universal values dataset actually yielded the lowest initial baseline scores at 0.460 for Gemma-3-4B-IT and 0.683 for Claude Sonnet 4.6. Additionally, in Track 2, the Modgenix submission scored slightly below the Claude baseline on the Indonesian dilemma dataset at 0.683 versus 0.696, and no team managed to surpass the high Claude baseline of 0.961 on the Sri Lankan societal values dataset. It is important to note that these accuracy variations cannot be attributed solely to

the cultural or reasoning difficulty of the scenarios, they also reflect substantial differences in dataset construction, annotation protocols, and how the target judgments are defined.

5.3 Effect of Fine-Tuning

In this shared task, we released 20% of the data as a development set with gold answers and 80% as a test set without gold answers. Despite the limited labeled data for model development, all participating teams were able to improve upon the baseline performance. This suggests that a relatively small amount of high-quality data can be effective in improving pluralistic value alignment in LLMs.

However, because the development and test sets share the same data distribution, these results primarily demonstrate strong alignment within this specific task. Further evaluation on out-of-distribution datasets will be necessary to confirm how well this data-efficient approach generalizes to unseen cultural scenarios.

5.4 Performance of the Recently Released LLMs

To observe model improvement, we refer to the methods used by the Modgenix team, which used GPT-5.4 with various prompting strategies, and the DUT-IR team, which used the recently released GPT-5.6 Sol to generate answers using country-specific guidance and few-shot examples. Modgenix achieved a score of 0.795, while DUT-IR achieved 0.803. Although using these recently released models demonstrates an increase in performance, the improvement is only marginal. From this, we can draw several insights, primarily that proprietary LLMs are paying greater attention to pluralistic value alignment. The continuous improvement of these models over successive releases highlights the growing importance of pluralistic value alignment in both research and industry.

6 Conclusion

In this paper, we have presented the findings of the PlurVA-LLM 2026 shared task on pluralistic value alignment across China, Indonesia, and Sri Lanka. In total we received 5 submissions from 4 teams. The methods proposed by the teams can surpass the baseline models in terms of macro average, highlighting the success of this shared task on pluralistic value alignment across 3 different countries and languages by using only a single LLM.

For future work, the next challenge is to further improve LLMs for pluralistic value alignment across countries and languages, particularly in multilingual settings. The relationship between countries and languages can be complex, where a country may have a single dominant language, a language may be shared across multiple countries, or a country may use multiple languages. It is essential to investigate how LLMs can effectively capture and align the pluralistic values of multiple countries while accounting for shared languages and multilingual contexts.

Limitations

We acknowledge several limitations in this shared task. First, our shared task focuses only on 3 countries: China, Indonesia, and Sri Lanka, with different dataset formats for each country. This limitation is mainly due to the limited availability of datasets specifically designed to evaluate pluralistic value alignment. Because of this, the concept of value alignment is interpreted differently across the three datasets. D2VBench selects options aligned with universal values, Pancasila-Dilemmas relies on the most frequent human response (majority vote), and LKValues evaluates the justifiability of specific statements. Consequently, a model matching the gold label does not carry the exact same meaning across these three cultural contexts.

Second, to keep the shared task simple and standardized, our evaluation relies primarily on exact-match accuracy against a single gold standard or consensus label, rather than using open-ended generation. We acknowledge that this simplified approach may fail to capture the natural diversity of opinions within each region.

Finally, due to time constraints following the workshop, we were unable to conduct a detailed, per-value error analysis to identify which specific value categories remain most challenging for current LLMs. We leave this diagnostic analysis for future work, and hope the released datasets will support further research and more extensive experiments.

Acknowledgments

This research was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0203000). We would like to thank the anonymous reviewers for their insightful comments.

References

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Carbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, and 80 others. 2026. Gemma 4 technical report. *CoRR*, abs/2607.02770.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular Pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

GLM. 2026. GLM-5: from vibe coding to agentic engineering. *CoRR*, abs/2602.15763.

Siyi Hao, Yidi Cao, Linhao Yu, Yuqi Ren, and Deyi Xiong. 2026. D2VBench: Benchmarking large language models with value dilemmas in daily scenarios. *Preprint*, arXiv:2607.19834.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew M. Bean, Katerina Margatina, Rafael Mosquera Gómez, Juan Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models.

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, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.

Nethmi Muthugala, Supryadi, Surangika Ranathunga, Nisansa de Silva, Ruijie Tao, Ovindu Gunatunga, Pengyun Zhu, Shaowei Zhang, Jingting Zheng, and Deyi Xiong. 2026. LKValues: Aligning large language models with Sri Lankan societal values. *Preprint*, arXiv:2607.20410.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. CIVICS: building a dataset for examining culturally-informed values in large language models. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24) - Full Archival Papers, October 21-23, 2024, San Jose, California, USA - Volume 1*, pages 1132–1144. AAAI Press.

Ashmari Pramodya, Nirasha Nelki, Heshan Shalinda, Chamila Liyanage, Yusuke Sakai, Randil Pushpananda, Ruvan Weerasinghe, Hidetaka Kamigaito, and Taro Watanabe. 2025. SinhalaMMLU: A comprehensive benchmark for evaluating multitask language understanding in Sinhala. *Preprint*, arXiv:2509.03162.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Alejandro Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan N. Gomez, Ivan Zhang, and 7 others. 2026. Tiny Aya: Bridging scale and multilingual depth. *CoRR*, abs/2603.11510.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2024a. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 19937–19947. AAAI Press.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024b. Position: A roadmap to pluralistic alignment. 235:46280–46302.

Supryadi, Irfan, Julianti, Darren Keanly Martin, Jayvin Fernando, Yuqi Ren, and Deyi Xiong. 2026. Pancasila-Dilemmas: Evaluating large language models on Indonesian human value dilemmas grounded in pancasila. *Preprint*, arXiv:2607.18066.

Gemma Team. 2025a. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Llama Team. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2025b. Qwen3 technical report. *CoRR*, abs/2505.09388.

Ping Wu, Guobin Shen, Dongcheng Zhao, Yuwei Wang, Yiting Dong, Yu Shi, Enmeng Lu, Feifei Zhao, and Yi Zeng. 2025. C-VARC: A large-scale Chinese value rule corpus for value alignment of large language models. *arXiv preprint arXiv:2506.01495*.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024. Exploring multilingual concepts of human values in large language models: Is value alignment consistent, transferable and controllable across languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1771–1793, Miami, Florida, USA. Association for Computational Linguistics.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. 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, Albuquerque, New Mexico. Association for Computational Linguistics.

Pardis Sadat Zahraei, Gokhan Tur, Dilek Hakkani-Tür, and Ehsaneddin Asgari. 2025. The alignment veto: How safety training suppresses cultural knowledge in llms. *arXiv preprint arXiv:2510.13154*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Jingting Zheng, Yuqi Ren, Linhao Yu, Yongqi Leng, and Deyi Xiong. 2026. Beyond value benchmarks: Measuring value-structure alignment in large language models via symmetric Q-sorts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18199–18230, San Diego, California, United States. Association for Computational Linguistics.

Pengyun Zhu, Yuqi Ren, Zhen Wang, Lei Yang, and Deyi Xiong. 2026. DVMap: Fine-grained pluralistic value alignment via high-consensus demographic-value mapping. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19834–19852, San Diego, California, United States. Association for Computational Linguistics.