PlurVA-LLM-2026 Shared Task Track-1: Pluralistic Value Alignment in LLMs via Multilingual Fine-Tuning and Threshold Calibration

Vihindi Kotalawala[♡] **& Nevidu Jayatilleke**[♢]
[♡]School of Computing, Informatics Institute of Technology, Sri Lanka
[♢]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
vihindi.20220672@iit.ac.lk, nevidu.25@cse.mrt.ac.lk

Abstract

We present our system for the *PlurVA-LLM 2026 Shared Task Track-1*, which focuses on pluralistic value alignment in the contexts of China, Indonesia, and Sri Lanka. For this resource-constrained track, we fine-tuned `Llama-3.1-8B-Instruct` using 4-bit QLoRA. Our approach combines option-permutation augmentation for Chinese data, annotator vote expansion for Indonesian data, and binary reformulation with SinhalaMMLU augmentation for Sri Lankan data. We further applied conditional threshold calibration to the predictions for the Sri Lankan data. The final system achieved accuracies of 0.785 for Chinese, 0.715 for Indonesian, and 0.916 for Sri Lankan, resulting in an overall macro-average accuracy of 0.805.

Code Model Dataset

1 Introduction

Large Language Models (LLMs) are increasingly used across diverse societies, yet conventional alignment often assumes a single set of human preferences. Human values vary across populations, motivating models that can represent diverse perspectives rather than a single universal preference (Sorensen et al., 2024). Existing alignment methods may instead learn an averaged preference and struggle to represent differences across cultures and communities (Feng et al., 2024). Multilingual capability alone also does not guarantee consistent value alignment, particularly across high- and low-resource languages (Xu et al., 2024).

The PlurVA-LLM 2026 Shared Task addresses this challenge by evaluating whether one system can make locally grounded value judgements across Chinese, Indonesian, and Sri Lankan contexts. The task combines three datasets with distinct cultural foundations and annotation structures. `D2VBench` presents value conflicts through daily-life scenarios from China (Hao et al., 2026). `Pancasila-Dilemmas` evaluates Indonesian judgements grounded in Pancasila and provides five citizen annotations for each question (Supryadi et al., 2026). `LKValues` evaluates societal values derived from Sri Lankan participants and expressed through Sinhala language (Muthugala et al., 2026).

We adopt a shared multilingual fine-tuning approach while adapting supervision to each dataset's characteristics. A shared adapter enables the model to learn across multiple cultural contexts, while country-specific prompts indicate the relevant value setting for each example (Xu et al., 2025). At the same time, dataset-specific preprocessing is used to account for differences in annotation formats, label structures, and observed preference distributions across the three datasets (Pezeshkpour and Hruschka, 2024; Sorensen et al., 2024). This allows us to retain dataset-specific value signals without requiring a completely separate model for each cultural context. For Sri Lankan data, where the supervision is comparatively limited and imbalanced, we further apply conditional threshold calibration at inference time to better map the model's statement-level confidence into the final four-way prediction (Esposito et al., 2021). Overall, our approach combines learning across all three datasets with specific adjustments for the distinct challenges present in each.

2 Related Work

Prior work on value alignment has increasingly considered pluralistic, multilingual, and culturally grounded preferences. Relevant studies span methods for representing heterogeneous human values, adapting models across cultural and linguistic contexts, and addressing challenges arising from preference disagreement, limited low-resource data, and cross-lingual variation. We review these directions below in relation to our multilingual value-alignment setting.

Proceedings of the First Workshop on Pluralistic Value Alignment of LLMs, pages 143–153
November 10, 2026 ©2026 Association for Computational Linguistics

2.1 Pluralistic Value Alignment

Pluralistic value alignment challenges the assumption that an LLM should optimise toward a single aggregate human preference. But as discussed in Section 1, relying mostly on a single reward signal can fail to represent diverse human preferences (Chakraborty et al., 2024), while diversified annotator preferences can also reduce the effectiveness of standard alignment methods (Zeng et al., 2024). There are three objectives for pluralistic alignment: Overton, steerable, and distributional pluralism, with distributional pluralism aiming to reflect the preference distribution of a community (Sorensen et al., 2024). Several approaches have consequently attempted to model such diversity explicitly. MaxMin-RLHF learns a mixture of reward models and optimises for diverse preference groups (Chakraborty et al., 2024), while Modular Pluralism combines a base LLM with specialised community models to support different value perspectives (Feng et al., 2024).

Value pluralism also extends across languages and societal contexts, where models may need to represent different locally grounded values rather than transfer a single dominant value system across populations. Existing LLMs can reflect the values of particular groups by default, motivating the development of methods specifically designed for cultural-value adaptation (Liu et al., 2025). CultureSPA addresses this through culture-joint and culture-specific fine-tuning, enabling a single model to serve multiple sets of culturally situated human values (Xu et al., 2025). Evaluation work has similarly moved beyond generic preference benchmarks. WorldValuesBench uses demographic contexts and real-world survey responses to evaluate whether language models reproduce diverse human-value distributions, and shows that this remains challenging even for strong models (Zhao et al., 2024).

2.2 Challenges in Value Alignment

Translating these pluralistic objectives into effective model training introduces several additional challenges. There is insufficient human preference data in many languages, making it harder to align models with the preferences expressed in those languages (Yang et al., 2025), while human-value representations can exhibit cross-lingual inconsistencies and asymmetric transfer between high- and low-resource languages (Xu et al., 2024). Recent Sri Lankan value-alignment work likewise reports persistent low-resource and cultural value-alignment gaps, with fine-tuning gains varying across model families (Muthugala et al., 2026). The structure of the supervision itself is also important. Preference datasets constructed using rigid annotation or aggregation schemes can suppress underlying preference heterogeneity (Chen et al., 2024). Standard preference-learning methods may not clearly distinguish between full agreement and a majority preference when annotators disagree (Zhang et al., 2024). Consequently, collapsing multiple annotations into a single majority label can remove meaningful minority judgments when disagreement reflects genuine value differences rather than annotation noise. Fine-tuning with such heterogeneous preferences is further affected by both model capacity and available data (Zeng et al., 2024). In multiple-choice value tasks, answer-option order can introduce an additional source of bias (Pezeshkpour and Hruschka, 2024).

Prior work has explored pluralistic preference learning, multilingual value alignment, and culturally grounded evaluation, often highlighting performance differences between high-resource and low-resource settings. In contrast, our setting requires a single model to perform well across all three datasets despite differences in language resources, annotation structures, and data availability. The goal is not only to measure these disparities but also to maximise performance across both high- and low-resource settings while avoiding improvements on one dataset at the expense of another. We address this by combining shared multilingual learning with dataset-specific supervision and inference strategies, aiming for consistently strong performance across the different value contexts and data conditions.

2.3 Language and Resource Context

The three communities involved in this shared task, Chinese, Indonesian, and Sri Lankan, are represented by the languages Chinese, Indonesian, and Sinhala. Each language corresponds to varying levels of data resources. Consequently, the strategy employed for each language is tailored to the availability of linguistic resources.

Chinese: Standard Chinese (Mandarin) is the official and predominant language of China and belongs to the Sino-Tibetan language family (Yang et al., 2022). It is written using Chinese characters

and is classified as a Category 5 (The Winners) language, representing a highly resourced NLP setting (Joshi et al., 2020; Ranathunga and de Silva, 2022). Among the three languages in this work, Chinese provides the highest resource capacity.

Indonesian: Indonesian (Bahasa Indonesia) is the national language of Indonesia and is widely used across its multilingual population (Gunarso et al., 2016). It has complex word-formation patterns and is spoken by more than 250 million people (Kamajaya and Moeljadi, 2025). Ranathunga and de Silva (2022) classify Indonesian as Category 4 (The Underdogs).

Sinhala: Sinhala is an Indo-Aryan language spoken primarily in Sri Lanka by approximately 16 million first-language speakers (Jayatilleke and de Silva, 2025). It uses a Brahmi-derived writing system (De Silva, 2026) and is classified as a lower-resourced Category 2 (The Hopefuls) language (Ranathunga and de Silva, 2022). Therefore, Sinhala is the most resource-constrained of the three languages considered in this work.

3 Methodology

The three PlurVA-LLM datasets differ substantially in their annotation and output structures. Chinese is a four-option multiple-choice task with one gold answer. Indonesian also contains four options, but each question is annotated by five individuals and may contain disagreement. Sri Lankan contains two candidate statements and requires one of four final labels: only statement A is correct, only statement B is correct, both are correct, or neither is correct.

We developed country-specific input and target transformations while retaining a single shared model architecture. All examples were converted into chat-formatted supervised fine-tuning records. The final system used `Meta Llama-3.1-8B-Instruct` as its base model and was adapted using 4-bit Quantised Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023). A single shared LoRA adapter was trained jointly on all 3 datasets: Chinese, Indonesian, and Sri Lankan.

All available labelled records were used for final training; no validation partition was reserved since the dataset was extremely small. The final training set contained 11,056 supervised records after country-specific transformations and augmentation shown in Figure 1, while the total evaluation set was 5475 (Chinese: 3,210, Indonesian:1,468, Sri Lankan:797).

Model	Chinese	Indonesian	Sinhala	Macro
`Qwen2.5-7B-Instruct`	0.4684	0.6351	0.4146	0.5060
`Aya Expanse 8B`	0.4494	**0.7027**	0.3415	0.4978
`Llama-SEA-LION-v3-8B-IT`	0.4620	0.6757	0.6829	0.6069
`Llama-3.1-8B-Instruct`	**0.5000**	0.6622	0.7561	**0.6394**
`Gemma 3n-E4B-it`	0.3608	0.6757	**0.8293**	0.6219

Table 1: Zero-shot model comparison.

Language	Experiment	Score
Chinese	− Option-permutation ensembling	0.779
	+ Option-permutation ensembling	**0.785**
Indonesian	Majority-vote hard-label training	0.651
	Annotator-vote expansion	0.689
	+ Option permutation	**0.715**
Sri Lankan	4-way classification (w/ adapter)†	0.832
	+ Binary formulation (w/ adapter)	**0.834**
	Binary formulation (w/o adapter)*	0.873
	+ MMLU augmentation	**0.890**
	+ Conditional calibration	**0.916**

† Language-specific adapter. * Multilingual shared adapter.

Table 2: Language-level experiments and selected configurations.

3.1 Base-model Selection

We first compared five instruction-tuned models under zero-shot prompting. Table 1 reports the macro average and country-level accuracies.

Although the strongest individual-language results varied across models. `Llama-3.1-8B-Instruct` for Chinese at 0.5, `Aya Expanse 8B` for Indonesian at 0.7027, and `Gemma 3n-E4B-it` for Sinhala at 0.8293 show the best-performing models for each language. But `Llama-3.1-8B-Instruct` achieved the highest macro-average of 0.6394, surpassing other models. It also provided a comparatively balanced performance across the three languages. Therefore, we selected it as the shared base model.

3.2 Country-specific Supervision

3.2.1 Chinese

The Chinese training set contained 790 questions. Each question included a value label, a question written in Chinese, four answer options, and one gold label from $\{A, B, C, D\}$. The original label distribution was imbalanced: 237 A, 123 B, 136 C, and 294 D.

The prompt instructions, the question, and the four answer options were written in Chinese. The model was prompted to return exactly one answer

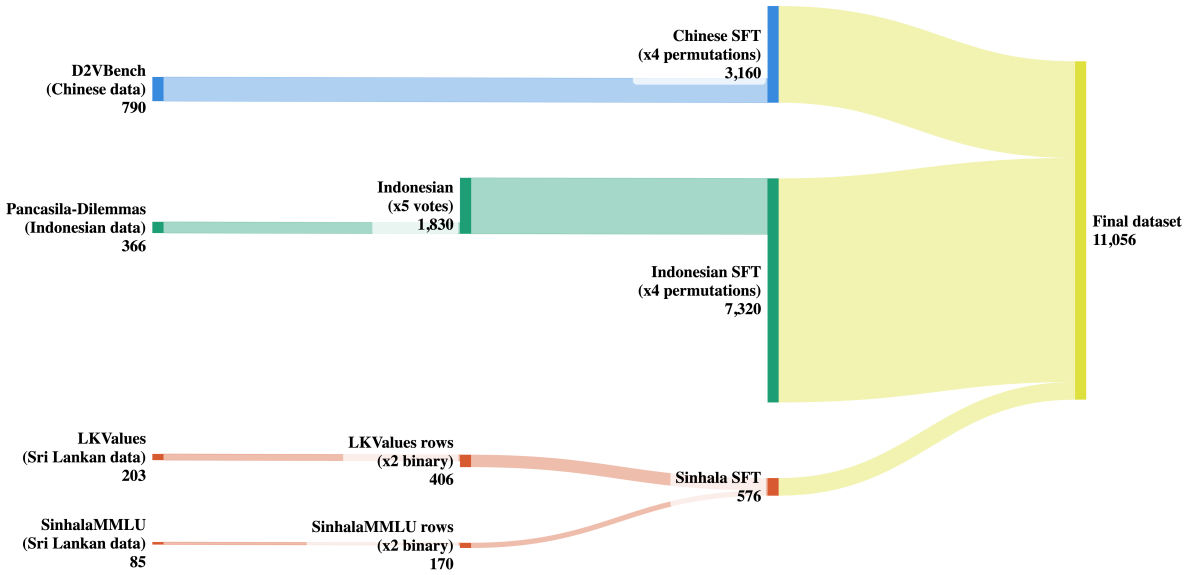


Figure 1: Data preparation and augmentation pipeline for the PlurVA-LLM training set. Flow widths are proportional to the number of training records produced at each stage for `D2VBench`, `Pancasila-Dilemmas`, `LKValues`, and `SinhalaMMLU` before merging into the final multilingual SFT dataset.

line in the form `Answer: X` as depicted in Figure 5a in Appendix A.

To reduce dependence on answer position, we applied cyclic option-permutation augmentation. For each question, four orderings of the options were constructed. When the options were shifted, the gold label was remapped to continue identifying the same answer content. As shown in Table 2, incorporating option permutation into the Chinese pipeline improved accuracy from 0.779 to 0.785.

The four cyclic orderings produced 3,160 Chinese SFT rows: $790 \times 4 = 3,160$. Because every original answer appeared once in each displayed position across the four cyclic permutations, the augmented target distribution contained exactly 790 instances of each output label. This simultaneously increased the amount of supervision and removed the original output-letter imbalance.

At inference, the same four cyclic orderings were scored. Each of the 3,210 original records was expanded into four permutations, resulting in a total of 12,840 records for prediction. Probabilities were mapped back to the original option identities and averaged before selecting the final answer. This made the final prediction less sensitive to the order in which options were displayed.

3.2.2 Indonesian

The Indonesian set contained 366 questions. Each question contained four answer options and five annotator votes, resulting in 1,830 individual annotations. The raw vote totals were 505 A, 402 B, 544 C, and 379 D. Of the 366 questions, 294 had a single consensus answer, while 72 had two answers in their consensus set. A majority-vote transformation would discard this disagreement and force each question to have a single hard target. In our experiments, as shown in Table 2, the majority-vote approach achieved an Indonesian accuracy of 0.651. Therefore, we retained all five annotations through vote expansion, thereby improving accuracy to 0.689.

The prompt instructions, the question, and the four options were written in Indonesian. The model was prompted to return only the answer in the form `Answer: X` as shown in Figure 5b in Appendix A.

For a question with annotator labels $y_1, \ldots, y_5$, we created one training record for every vote. For each Indonesian question, the five annotator votes were retained as separate training targets. Their contributions were represented through the log-probabilities of the corresponding answers, so answers receiving more votes contributed proportionally more to the learning objective while minority judgments were still preserved. This allowed the empirical annotator distribution to be learned without introducing a custom soft-label loss.

We additionally applied the same four cyclic option permutations used for Chinese. Each Indonesian question consequently generated (5 votes $\times$

4 permutations = 20) training records, producing 7,320 Indonesian SFT rows, which increased the accuracy from 0.689 to 0.715, as shown in Table 2.

During inference, we employed the same strategy used for the Chinese evaluation dataset. Each of the 1,468 original records was expanded into four permutations, resulting in a total of 5,872 records for prediction. The predicted probabilities across the four permutations were then averaged to determine the final answer for each original record.

3.2.3 Sri Lankan

The labelled Sri Lankan training set contained 203 questions written entirely in Sinhala. Its original four-way label distribution was 107 A, 33 B, 43 Both, and 20 neither-correct. This direct four-way training would expose the model to a considerable output-class imbalance.

Since the two statements were independent in the dataset, we decomposed every question into two independent binary decisions:

$$z_A = \mathbb{I}[\text{statement A is correct}]$$
$$z_B = \mathbb{I}[\text{statement B is correct}]$$

Each question generated two SFT records: one presenting statement A and one presenting statement B. The target for each record was either `Answer: Yes` or `Answer: No`. At inference, the two decisions were recombined as follows:

$$(z_A, z_B) = \begin{cases} (1,0) & \to A, \\ (0,1) & \to B, \\ (1,1) & \to \text{Both}, \\ (0,0) & \to 0. \end{cases}$$

The 203 original questions produced 406 binary records, consisting of 226 Yes targets and 180 No targets. This reformulation reduced the severity of the original four-class imbalance while preserving all four possible final outcomes. During the multi-adapter experiment, in which separate language-specific adapters were trained, we compared the four-way and binary formulations for Sinhala. As shown in Table 2, the binary formulation performed slightly better, improving accuracy from 0.832 to 0.834.

The prompt retained the original Sinhala question and candidate statement. The system and task instructions were written in English and described the broader Sri Lankan context, including the Sinhalese, Tamil, Muslim, and Burgher communities. The corresponding Sri Lankan value label was also included. The model was prompted to return only `Answer: Yes` or `Answer: No` as shown in Figure 5c in Appendix A.

During inference, each record was split into two separate questions, following the same approach used during training. As a result, the original 797 records were expanded into 1,594 records for prediction.

3.2.3.1 SinhalaMMLU Augmentation

The Sri Lankan dataset had substantially fewer training examples than the Chinese and Indonesian datasets. We augmented it with manually selected questions from `SinhalaMMLU` (Pramodya et al., 2025). From the original resource, we selected 68 Civics questions and 17 Health and Physical Science questions, giving 85 additional questions.

Selection was performed manually according to relevance to the shared task. Each selected MMLU question originally contained four answer options and one correct option. We converted these records into the same two-statement Sinhala format used by the main dataset.

Two conversion strategies were used:

1. When the original question could be retained, the correct answer and the most plausible alternative among the remaining answers were selected as the two candidate statements.

2. When a meaningful two-statement record could not be produced directly, the question and candidate statements were manually amended or rewritten while preserving the relevant knowledge and value judgment by the authors of this study, who are native Sinhala speakers. Figure 2 presents an example in which the answer choices of a SinhalaMMLU Health and Physical Science question were manually adapted into two candidate statements.

Each converted question contained one correct and one incorrect statement, producing 85 Yes and 85 No records. The MMLU augmentation[1] contributed 170 balanced binary records.

`Gemini 3.1 Flash` was used only to assign one value-category label to each converted record. All

[1] 🤗 PlurVA-LLM-Sinhala-Augmented

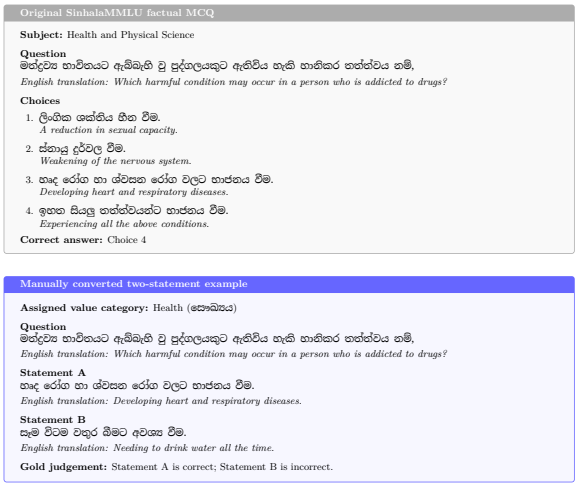


Figure 2: Example of manually adapting the answers of a SinhalaMMLU Health and Physical Science question into two candidate statements.

generated value labels were manually refined by the authors, who are native Sinhala speakers, before inclusion.

Combining the main Sinhala data with MMLU augmentation yielded 288 questions and 576 binary SFT rows: $(203 + 85) \times 2 = 576$. The combined binary distribution contained 311 Yes and 265 No targets. As shown in Table 2, this augmentation improved accuracy from 0.873 to 0.890.

3.3 Prompt and Target Design

All three countries' prompts explicitly included the supplied value category. Chinese and Indonesian used instructions in their respective languages, while Sri Lankan retained the original Sinhala content with concise English task instructions. Prompts for the respective langagues are mentioned in Appendix A.

During development, we additionally experimented with rationale supervision and value-description summaries. These configurations and their corresponding performance are reported in Appendix B. However, neither rationales nor value summaries were included in the final selected model because they did not improve the overall performance.

This result suggests that the extra explanations added unnecessary or noisy information for the classification task. Zhu et al. (2025) has similarly shown that adding rationales does not always improve language-model performance and can sometimes reduce accuracy. More generally, adding extra training signals can lead to negative transfer when they introduce task-irrelevant information or misleading patterns. Since the required outputs in this task were short categorical decisions, directly supervising the target labels provided a more focused learning signal.

Hyperparameter	Final value
Base model	`Meta Llama-3.1-8B-Instruct`
Quantisation	4-bit NF4 with double quantization
Compute data type	bfloat16
Epochs	2
Peak learning rate	1×10^{-4}
Learning-rate schedule	Cosine decay
Warm-up ratio	0.05
Per-device batch size	2
Gradient accumulation	10
Effective batch size	20
Maximum sequence length	1,536
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
Maximum gradient norm	1.0
Weight decay	0.0
Random seed	42

Table 3: Hyperparameters of the selected shared adapter.

Every assistant target in the final model consisted only of a short machine-readable answer marker: "`Answer: A/B/C/D`" for Chinese and Indonesian, or "`Answer: Yes/No`" for Sri Lankan.

3.4 QLoRA Fine-tuning

The model was fine-tuned using QLoRA. Trainable LoRA matrices were inserted into seven attention and feed-forward projections:

$$\{\texttt{q_proj}, \texttt{k_proj}, \texttt{v_proj}, \texttt{o_proj}, \texttt{gate_proj}, \texttt{up_proj}, \texttt{down_proj}\}.$$

Training used a standard causal language model cross-entropy with the prompt tokens masked, so the loss was computed only over the assistant's response. Since the targets contained only the final answer markers, the optimisation focused directly on the required output labels.

We compared separate language-specific adapters with our proposed shared multilingual adapter, jointly trained on all three languages. As shown in Table 4, the shared adapter achieved the highest macro-average accuracy of 0.796, compared with 0.770 for the language-specific setup, and was selected as our final approach. The shared setup was particularly effective for Indonesian and Sri Lankan data, reaching 0.715 and 0.89, respectively. This is consistent with Xu et al. (2025), who report that joint fine-tuning can outperform culture-specific tuning by enabling knowledge sharing across groups. The final

Adapter approach	Chinese	Indonesian	Sinhala	Macro avg
Language-specific adapters	**0.788**	0.687	0.834	0.770
Our approach (shared adapter)	0.785	0.715	0.89	0.796
+ threshold calibration	0.785	**0.715**	**0.916**	**0.805**

Table 4: Comparison of multiple language-specific adapters and a shared multilingual adapter

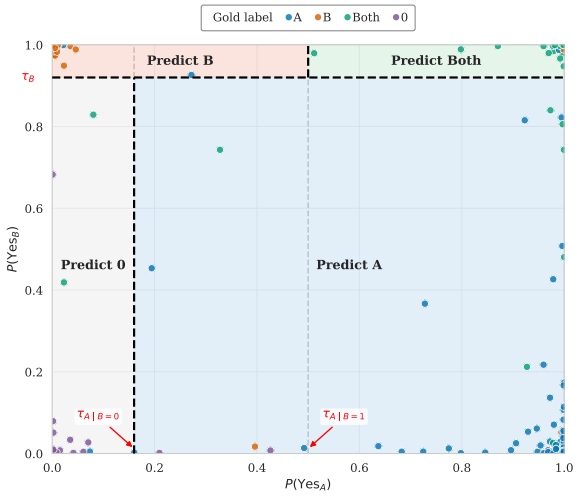


Figure 3: Conditional decision boundaries mapping Sinhala statement-level probabilities to the final classes A, B, Both, and 0.

hyperparameters used for the best model are depicted in Table 3.

3.4.1 Conditional Threshold Calibration

After fine-tuning, we calibrated the Sri Lankan decision thresholds using five-fold cross-validation on the training set. Because the original four-way prediction was reconstructed from two binary statement-level decisions, applying a fixed threshold of 0.5 to both statements did not provide the best class balance. Esposito et al. (2021) has similarly shown that fixed decision thresholds can be suboptimal under class imbalance and that post hoc threshold adjustment can improve classification performance. In particular, the model tended to over-predict cases in which statement B was considered valid. We used the five folds to identify suitable directions and ranges for adjusting the conditional thresholds.

The final prediction was determined using three thresholds. We first evaluated statement B using τ_B. The threshold for statement A was then selected conditionally: $\tau_{A|B=0}$ was applied when B was predicted as invalid, while $\tau_{A|B=1}$ was applied when B was predicted as valid. The resulting binary decisions were then mapped back to the original four classes: A, B, Both, and 0.

Exp.	τ_B	$\tau_{A\|B=0}$	$\tau_{A\|B=1}$	Accuracy
E0	0.50	0.50	0.50	0.8900
E1	0.70	0.40	0.50	0.9100
E2	0.80	0.30	0.50	0.9100
E3	0.85	0.25	0.50	0.9100
E4	0.90	0.20	0.50	0.9133
E5	0.92	0.16	0.50	**0.916**

Table 5: Conditional threshold calibration for the Sinhala statement-level predictions.

Figure 3 visualises the resulting decision regions. The horizontal boundary τ_B separates cases where B is predicted as valid from those where it is not. Below this boundary, $\tau_{A|B=0}$ separates predictions 0 and A, whereas above it, $\tau_{A|B=1}$ separates B and Both. Thus, the threshold applied to statement A depends on the preceding decision for statement B. A further detailed threshold sensitivity analysis is presented in Appendix C.

Based on the five-fold cross-validation results, we narrowed the search to a smaller set of candidate threshold combinations. These configurations are reported in Table 5, beginning with the default 0.5 thresholds and progressively adjusting τ_B and $\tau_{A|B=0}$ while keeping $\tau_{A|B=1} = 0.50$.

The progression in Table 5 shows that increasing τ_B while decreasing $\tau_{A|B=0}$ improved performance over the default configuration. The selected setting used $\tau_B = 0.92$, $\tau_{A|B=0} = 0.16$, and $\tau_{A|B=1} = 0.50$, improving accuracy from 0.8900 with the default thresholds to 0.916. As shown in Table 4, applying this calibration to Sri Lankan data with our shared multilingual adapter produced the final selected model, achieving the highest macro-average accuracy of 0.805.

The complete prediction procedure using the selected thresholds is illustrated in Figure 4. For each question, the model first produces separate Yes/No probabilities for statements A and B. Statement B is evaluated using $\tau_B = 0.92$. If B is predicted as invalid, statement A is evaluated using $\tau_{A|B=0} = 0.16$; otherwise, $\tau_{A|B=1} = 0.50$ is used. The resulting binary pair (z_A, z_B) is then recombined into the final four-way prediction: $(1,0) \rightarrow$ A, $(0,1) \rightarrow$ B, $(1,1) \rightarrow$ Both, and $(0,0) \rightarrow 0$.

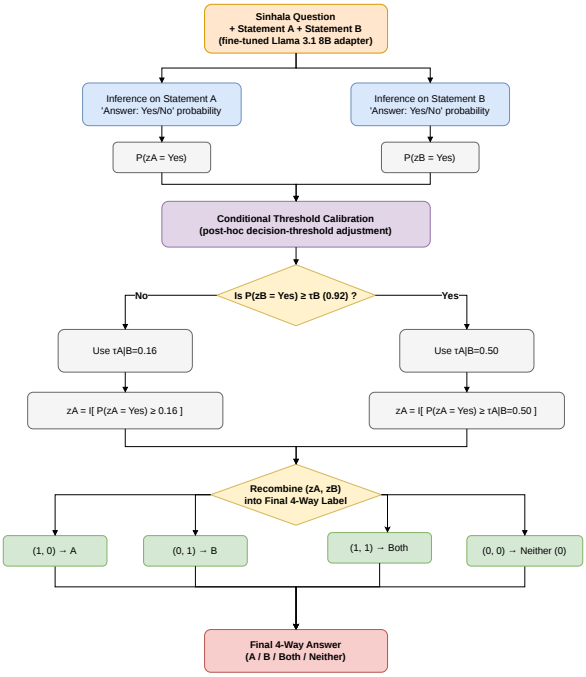


Figure 4: Conditional threshold calibration and four-way reconstruction for Sinhala predictions.

Although $\tau_B = 0.92$ is numerically high, inspection of the predicted probability distribution showed that relatively few instances fell below this threshold. Therefore, it primarily filtered lower-confidence B decisions that contributed to over-prediction rather than suppressing confident B predictions. Overall, the results suggest that asymmetric conditional thresholds were better suited to the model's probability distribution than applying a uniform threshold of 0.5 to both statement-level decisions.

4 Conclusion

We presented our system for the *PlurVA-LLM 2026 Shared Task Track-1* (the resource-constrained track), addressing pluralistic value alignment across Chinese, Indonesian, and Sri Lankan. Our approach uses a shared `Llama-3.1-8B-Instruct` model while applying dataset-specific supervision strategies, including cyclic option permutation for Chinese, annotator vote expansion for Indonesian, and binary reformulation with `SinhalaMMLU` augmentation and conditional threshold calibration for Sri Lankan data. Our experiments highlight the challenges of multilingual value alignment when datasets differ in language resources, annotation structures, label distributions, and the amount of available labelled data.

Important challenges remain, including preserving diverse annotator opinions, working with limited value-alignment data, reducing answer-position bias, and adapting a shared multilingual model to datasets with different characteristics. Future work should investigate larger and more diverse value-alignment datasets, improved methods for modeling annotator disagreement, and more robust multilingual training strategies that can transfer value knowledge across languages without reducing performance in individual datasets. Overall, our results suggest that combining shared multilingual learning with supervision tailored to each dataset's characteristics is a promising direction for pluralistic value alignment.

Limitations

During the implementation of the solution, we encountered several limitations.

Data Scarcity: The training data are limited and unevenly distributed across the three languages, which may affect how well the shared model learns each value context. Since all labelled data were used for training, no independent validation set was available.

Model Size Limit: Following the constraints of the *PlurVA-LLM Shared Task Track-1*, our experiments were limited to models with up to 8B parameters. Therefore, the effectiveness of the proposed strategies may differ for larger models or other model families.

Asymmetric Calibration: The asymmetric Sinhala calibration may be sensitive to statement order, as statement B is evaluated before selecting the conditional threshold for statement A.

References

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. 2024. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*.

Nisansa De Silva. 2026. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Carmen Esposito, Gregory A Landrum, Nadine Schneider, Nikolaus Stiefl, and Sereina Riniker. 2021. GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling*, 61(6):2623–2640.

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.

Gunarso, Hammam Riza, Elvira Nurfadhilah, M. Teduh Uliniansyah, Agung Santosa, and Lyla R. Aini. 2016. An Overview of BPPT's Indonesian Language Resources. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 73–77, Osaka, Japan. The COLING 2016 Organizing Committee.

Siyi Hao, Yidi Cao, Linhao Yu, Yuqi Ren, and Deyi Xiong. 2026. D2VBench: Benchmarking Large Language Models with Value Dilemmas in Daily Scenarios. *arXiv preprint arXiv:2607.19834*.

Nevidu Jayatilleke and Nisansa de Silva. 2025. SiDiaC: Sinhala Diachronic Corpus. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 511–527, Hanoi, Vietnam. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ian Kamajaya and David Moeljadi. 2025. IndoMorph: a morphology engine for Indonesian. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 72–81, Online. Association for Computational Linguistics.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-Instruction: Aligning Language Model in Low-resource Languages with Self-curated Cross-lingual Instructions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 546–566, Bangkok, Thailand. Association for Computational Linguistics.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025. Cultural Learning-Based Culture Adaptation of Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Nethmi Muthugala, 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. *arXiv preprint arXiv:2607.20410*.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32943–32961, Suzhou, China. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

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. *arXiv preprint arXiv:2607.18066*.

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.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025. Implicit Cross-Lingual Rewarding for Efficient Multilingual Preference Alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21125–21147, Vienna, Austria. Association for Computational Linguistics.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese Minority Pre-trained Language Model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dun Zeng, Yong Dai, Pengyu Cheng, Longyue Wang, Tianhao Hu, Wanshun Chen, Nan Du, and Zenglin Xu. 2024. On Diversified Preferences of Large Language Model Alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9194–9210, Miami, Florida, USA. Association for Computational Linguistics.

Michael JQ Zhang, Zhilin Wang, Jena D Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2024. Diverging preferences: When do annotators disagree and do models know? *arXiv preprint arXiv:2410.14632*.

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.

Chiwei Zhu, Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, and Zhendong Mao. 2025. Rationales Are Not Silver Bullets: Measuring the Impact of Rationales on Model Performance and Reliability. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5808–5835, Vienna, Austria. Association for Computational Linguistics.

A Prompt Templates

This section presents the complete prompt templates used across the three language-specific experiments. Each prompt contains a system instruction, a user template, and a constrained output format. Placeholders enclosed in braces were replaced with the corresponding fields from each dataset, such

```
[SYSTEM]
你是一位深谙中国社会价值观与日常伦理困境的助手。题目描述了一个第一人称的价值冲突情境，四个选项都是现实中可能的做法。你的任务是选出最能践行指定核心价值、同时符合中国社会主流判断的选项。

[USER]
本题考察的核心价值：{value_native}（{value_english}）

情境与问题：
{question}

选项：
A. {opt_a}
B. {opt_b}
C. {opt_c}
D. {opt_d}

请直接给出最终选择，只输出一行，格式为 “Answer: X”（X 为 A/B/C/D）。
```

(a) Chinese prompt

```
[SYSTEM]
Anda adalah asisten yang memahami nilai-nilai Pancasila dan norma masyarakat Indonesia. Tugas Anda adalah memilih tindakan yang paling sesuai dengan penilaian mayoritas masyarakat Indonesia terhadap nilai yang diuji.

[USER]
Nilai yang diuji: {value_native}

Skenario:
{scenario}

Pertanyaan:
{question}

Pilihan:
A. {opt_a}
B. {opt_b}
C. {opt_c}
D. {opt_d}

Berikan jawaban akhir Anda dalam satu baris saja dengan format "Answer: X" (X adalah A/B/C/D).
```

(b) Indonesian prompt

```
[SYSTEM]
You are an assistant with deep familiarity with Sri Lankan societal values across its Sinhalese, Tamil, Muslim and Burgher communities. You will read a question and one candidate statement, both in Sinhala. Judge whether the statement is a correct and socially appropriate answer to the question according to the given Sri Lankan value.

[USER]
Value being tested: {value_english}

Question (Sinhala):
{question}

Candidate statement (Sinhala):
{statement}

Reply with exactly one line: "Answer: Yes" if the statement is a correct and appropriate answer to the question, or "Answer: No" if it is not.
```

(c) Sri Lankan prompt

Figure 5: Prompt templates used for the Chinese, Indonesian, and Sri Lankan datasets.

as the value label, question, scenario, or candidate statements. The 3 different templates for each dataset are shown in Figure 5.

For Chinese and Indonesian, both the task instructions and input content were presented in the respective languages. For Sinhala, the question and candidate statements were kept in Sinhala, while the task instructions were written in English. This cross-lingual prompt design follows Li et al. (2024), showing that English instructions can be effectively combined with low-resource-language content, where instruction-following capability may be stronger in English. In all cases, the prompts were constrained to produce short, machine-readable outputs matching the target format.

B Negative Experiments

We investigated whether rationale supervision and contextual information at the value level improved multilingual value alignment. The same model architecture and training hyperparameters were used in all four experiments. The only difference was whether generated rationales and value summaries were included.

Experiment	Macro Avg.
Direct-answer supervision	**0.776**
Value-summary context	0.771
Rationale supervision	0.759
Rationale supervision with value-summary context	0.759

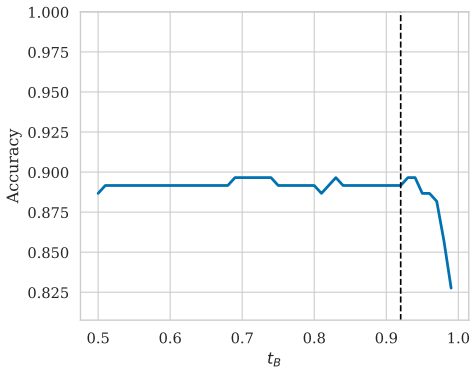
Table 6: Performance of rationale and value-summary experiments.

Rationales are short explanations of why a selected answer is appropriate for the value being tested. They were generated using `Llama-3.1-8B-Instruct`, and only rationales that produced the expected final answer were retained.

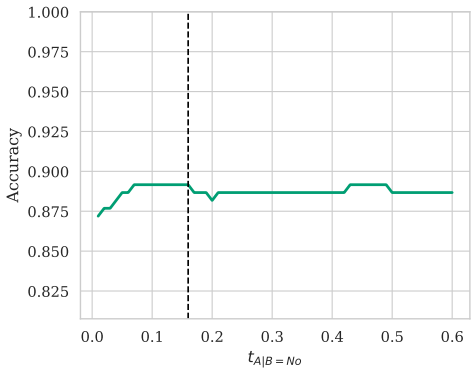
Value summaries are general descriptions of how a particular value is applied across multiple examples. They were created using the same model by grouping the generated rationales under each value and summarising their common reasoning patterns. The performance of rationale and value-summary experiments is presented in Table 6.

C Threshold Sensitivity Analysis

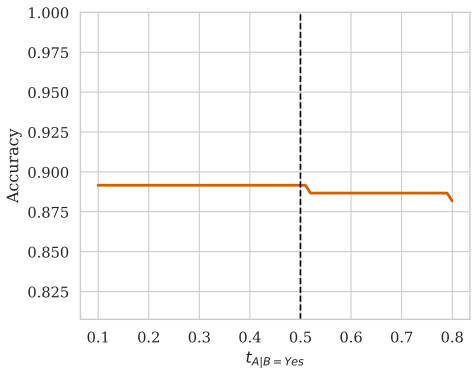
During threshold calibration, we first varied τ_B and evaluated statement-B accuracy at each threshold. We selected $\tau_B = 0.92$ because it produced the highest statement-B accuracy. We then fixed τ_B at 0.92 and calibrated the threshold for statement A under the two possible predictions for statement B. When B was predicted as No, the selected threshold was $\tau_{A|B=0} = 0.16$. When B was predicted as Yes, the selected threshold was $\tau_{A|B=1} = 0.50$. The final threshold configuration was therefore $(\tau_B, \tau_{A|B=0}, \tau_{A|B=1}) = (0.92, 0.16, 0.50)$.



(a) Sensitivity to t_B



(b) Sensitivity to t_A when $B = \text{No}$



(c) Sensitivity to t_A when $B = \text{Yes}$

Figure 6: Effect of conditional threshold values on out-of-fold Sinhala accuracy.