PlurVA-LLM-2026 Shared Task Track-1: Curriculum Fine-Tuning with Soft Labels and Positional Consistency for Cross-Cultural Value Alignment

Amasha Widanagamage[1], Hansika Seneviratne[1], Naween Fernando[2], Dilushri Pavithra[1], Ruvan Weerasinghe[1], Deshan Sumanathilaka[3]

[1]Informatics Institute of Technology, Sri Lanka, [2]University of Moratuwa, Sri Lanka, [3]Swansea University, Wales, UK

Correspondence: hansika.s@iit.ac.lk

Abstract

Pluralistic value alignment asks whether a language model can adapt its judgements to distinct, locally grounded value systems rather than reflecting a single aggregated notion of human preference. The PlurVA-LLM shared task evaluates this setting by unifying three independently constructed, culturally grounded multiple-choice value dilemmas across Chinese, Indonesian, and Sri Lankan datasets. We present a single-model submission based on Qwen3.5-4B with LoRA fine-tuning, trained only on the 1,359 labelled development instances provided by the task. Our method combines three ideas motivated directly by the task structure: soft-label supervision for Indonesian annotator-vote distributions, consistency regularisation across option permutations to reduce answer-position bias, and a staged multilingual curriculum with cross-validation-based iteration selection. On the hidden test set, our system achieves a macro-average accuracy of 0.8062, with 0.7788 on Chinese, 0.7139 on Indonesian, and 0.9260 on Sri Lankan, outperforming a manually tuned reference curriculum from our internal development. These results show that a single parameter-efficient model can unify culturally and structurally distinct value-alignment benchmarks under a low-resource training budget.

1 Introduction

The rapid growth of *Large Language Models* (LLMs) as globally deployed digital infrastructure has made value alignment, the problem of ensuring model outputs reflect appropriate human values, a central concern in AI research. Standard alignment pipelines optimise towards an aggregated human preference, which in practice approximates an "average" annotator and marginalises populations whose values differ from that average (Sorensen et al., 2024). Empirically, LLM value expressions have repeatedly been found to cluster near Western, English-speaking populations (Hershcovich et al., 2022; Santurkar et al., 2023; Durmus et al., 2024). The PlurVA-LLM Shared Task asks whether a model can instead adapt its judgements to distinct, locally grounded value systems. Given a scenario and a target country, a system selects the response most consistent with that country's annotated human value preferences, across Chinese (Hao et al., 2026), Indonesian (Supryadi et al., 2026) and Sri Lankan (Muthugala et al., 2026) data, and is scored on the macro average over all three. Track-1 additionally constrains systems to a single model of at most 8B parameters, with no external APIs, agents, tools or knowledge bases at inference time. [1]

2 Related Works

2.1 Multilingual and Low-Resource Benchmarks

Multilingual question answering and evaluation resources have expanded rapidly across languages (Perevalov et al., 2024), and establish the setting in which our three subsets sit. `TyDi QA` (Clark et al., 2020) provides information-seeking QA written directly in eleven typologically diverse languages, including Indonesian, and `INCLUDE` (Romanou et al., 2024) evaluates regional knowledge through multiple-choice items spanning many languages including Sinhala. For Sinhala specifically, resources remain scarce. Early work built classifier-based QA over a news corpus (Jayakody et al., 2016), `SiQuAD` (Ranasinghe and Weerasinghe, 2025) supplies span-based QA by translating and cleaning `SQuAD v1.1`, curriculum-grounded retrieval systems have been proposed for the Sri Lankan school context (Kiridana and Dias, 2025), and `SinhalaMMLU` (Pramodya et al., 2025) provides multitask multiple-choice evaluation. These benchmarks measure factual and task competence rather than value alignment, and none of them models

[1]Team SLNLP. Code available at: `https://github.com/Schizo00/PlurvaLLM-SharedTask`

Proceedings of the First Workshop on Pluralistic Value Alignment of LLMs, pages 187–196
November 10, 2026 ©2026 Association for Computational Linguistics

the annotator disagreement or the locally grounded normative judgements the present task requires.

2.2 Pluralistic and Cross-Cultural Value Alignment

Standard *Reinforcement Learning from Human Feedback* (RLHF) and instruction-tuning paradigms tend to aggregate preferences into a single "average human" response, effectively marginalising diverse socio-cultural and regional perspectives. Sorensen et al. (2024) formalise three targets for pluralistic alignment (overton, steerable and distributional) and argue that a single aggregated preference model satisfies none of them. Empirically, model value expressions cluster near Western, English-speaking respondents (Durmus et al., 2024; Santurkar et al., 2023), and cross-cultural evaluations find that models default to Western normative framings even when prompted in another language (AlKhamissi et al., 2024; Naous et al., 2024). `PRISM` (Kirk et al., 2024) addresses this on the data side, collecting preferences from participants across many countries and retaining rater identity rather than collapsing it. The PlurVA-LLM benchmark takes a complementary position: rather than a single multi-country preference corpus, it composes three independently constructed, natively formatted resources and requires one system to serve all of them. To address the aggregation limitation, recent research has transitioned towards pluralistic alignment models that support multi-perspective steering.

Feng et al. (2024) introduce Modular Pluralism, a multi-LLM collaboration framework that plugs specialised, community-tailored LMs into a central base LLM. Their framework enables models to support overton, steerable, and distributional pluralism without modifying black-box weights. Expanding on this, Zhu et al. (2026) argue that national-level labels are too coarse to capture intra-country preference variance. They propose DVMap (High-Consensus Demographic-Value Mapping), which leverages multi-dimensional demographic attributes from the *World Values Survey* (WVS), structured *Chain-of-Thought* (CoT) reasoning, and *Group Relative Policy Optimization* (GRPO) to achieve fine-grained pluralistic alignment across cross-demographic and cross-cultural evaluation splits.

2.3 Learning from Annotator Disagreement

Traditional supervised NLP pipelines treat annotator disagreement as noise, collapsing varying judgements into a single majority-vote label. However, emerging paradigms recognise disagreement as a valuable signal reflecting linguistic ambiguity, task complexity, or personal perspective.

Early foundational work by Fornaciari et al. (2021) demonstrates that soft-label probability distributions can be leveraged via multi-task auxiliary heads to mitigate overfitting and improve generalisation on inherently ambiguous tasks. Building upon this, Xu et al. (2024) evaluate disagreement-aware strategies across standard text classification benchmarks, illustrating that directly predicting soft annotator distributions yields superior performance over single-label classifiers. More recently, Benito-Santos and Ghajari (2025) extend this principle to *Semantic Textual Similarity* (STS), showing that modelling full ordinal annotator distributions via distance-aware cross-encoders and truncated Gaussian heads significantly enhances calibration and ranking accuracy on ambiguous sentence pairs.

2.4 Selection Bias in Multiple-Choice Questions

Multiple-Choice Question Answering (MCQA) is a primary evaluation standard for LLMs and *vision-language models* (VLMs). However, models frequently exploit spurious positional priors or symbol biases rather than demonstrating genuine reasoning. Zheng et al. (2024) show that LLMs carry a consistent bias towards particular option identifiers, that this bias is a substantial source of benchmark variance, and that debiasing by permutation materially changes measured accuracy; Pezeshkpour and Hruschka (2024) report the same sensitivity to option ordering across several model families.

For multimodal models, Loginova et al. (2025) analyse position-based selection bias in Video-LLMs and introduce BOLD, a calibration technique designed to suppress "blind guessing" across complex video reasoning tasks. Similarly, Guda et al. (2025) analyse selection bias in text-only LLMs. They formulate the *Permutation Bias Metric* (PBM), an unsupervised, label-free metric that quantifies prediction instability across option permutations, and present BaQCKV (Batch Question-Context KV caching) alongside a LoRA fine-tuning strategy to eliminate positional bias efficiently.

2.5 Consistency Regularisation

Consistency regularisation enforces model invariance under semantics-preserving input perturbations, serving as an effective inductive bias to prevent models from relying on superficial cues.

In the realm of structural reasoning, Yin et al. (2023) apply consistency regularisation training to Transformers for compositional generalisation. Their approach enforces both representation consistency across distinct inputs and prediction consistency under equivalent rephrasings, demonstrating marked improvements in semantic parsing and machine translation. In evaluation scenarios, consistency regularisation such as the objective studied in Guda et al. (2025) is applied directly across option permutations to stabilise LLM outputs against positional shifts, bridging the gap between raw likelihood estimation and robust multiple-choice reasoning.

2.6 Low-Resource Languages and Efficient Adaptation

Data scarcity and computational cost continue to limit advanced NLP for low-resource and underrepresented languages.

For Marathi, Khade et al. (2025) evaluate *Low-Rank Adaptation* (LoRA) as a form of *Parameter-Efficient Fine-Tuning* (PEFT). They observe that target-language generation improves, but standard automated metrics often misalign with human evaluations and mask trade-offs in general reasoning capabilities. To maintain general instruction-following while acquiring new languages, Sarasua et al. (2025) propose DIPLomA. This decoupled instruction-preserving language-adaptation method merges post-training deltas and separates target-language continual pre-training from instruction alignment. Finally, Joshi et al. (2025) show that continued pre-training using a mixture of real and synthetic translation corpora (e.g., Nemotron-Mini-Hindi 4B) significantly improves factual accuracy and task performance in low-resource Indic languages.

3 Methodology

The proposed system is developed through a series of systematic steps to unify three typologically distinct value-alignment datasets (Chinese, Indonesian, and Sri Lankan) into a single fine-tuned model capable of producing culturally grounded, contextually appropriate judgements across all three, despite differences in native task format, annotation structure, and language resource availability.

3.1 Dataset Analysis

The benchmark combines three independently constructed resources: Chinese daily-life value dilemmas from `D2VBench` (Hao et al., 2026), Indonesian dilemmas grounded in Pancasila values (Supryadi et al., 2026), and Sri Lankan value-judgement items from LKValues (Muthugala et al., 2026). Table 1 summarises the resulting dev/test splits answer spaces.

Dataset	Dev	Test	Answer space
Chinese	790	3,210	A/B/C/D
Indonesian	366	1,468	A/B/C/D
Sri Lankan	203	797	A/B/Both/0
Total	1,359	5,475	

Table 1: Dataset details. Each resource is split 20/80; only the development portion carries gold labels, and the test portion is scored on Codabench.

The subsets differ in how values are grounded as much as in format. `D2VBench` builds first-person dilemmas through LLM-human collaboration around 158 fine-grained value concepts, pairing internal conflict with external constraint. Every option is plausible, including compromise and retreat rather than only normatively correct choices. Pancasila-Dilemmas derives scenarios from Indonesian news articles classified under the five Pancasila values, with correctness set by five-annotator consensus rather than a single gold label. LKValues grounds its 40 societal values in a trilingual survey of 205 respondents spanning the Sinhalese, Tamil, Muslim and Burgher communities rather than adopting an external framework, and combines `SinhalaMMLU` reformulations (Pramodya et al., 2025) with generated scenarios validated by annotators from each group. Its items present a question stem with two candidate statements and admit A, B, `Both`, or `0` (neither), where the latter two account for 31% of development rows.

Label structure: As our method is motivated by the annotation rather than by architecture, we report the statistics that drove the design. Gold letters are unevenly distributed: in Chinese, A-237, B-123, C-136, D-294; in Sri Lankan, A-106, B-34, Both-43, 0-20. The most frequent Chinese letter (D) is $2.4\times$ more frequent than the least (B), and the most frequent Sri Lankan label (A) is $5.3\times$ more

frequent than the least (0). There is no reason to expect these priors to transfer to the hidden test set.

The Indonesian annotation is distributional rather than categorical. All 366 development rows carry five votes; only 26 rows (7.1%) are unanimous, 103 rows have a top vote of just two out of five, and 72 rows (19.7%) have no strict majority at all.

Evaluation: Systems submit one prediction file covering all three countries, scored by per-country accuracy and their macro average. For Indonesian, the reference is the consensus of the annotator votes. The answer or answers with the highest counts and where a tie occurs all tied answers are accepted.

This evaluation design has direct implications for system design. A model trained on only majority labels would implicitly ignore that up to one-fifth of Indonesian annotators disagree with the stated consensus, while a system relying on label frequency risks inheriting label imbalances from the Chinese and Sri Lankan datasets that may not hold on the hidden test set. These two properties, annotator disagreement in the Indonesian data and unstable label priors across all three datasets, directly motivate two design choices in our system, treating Indonesian gold as a soft distribution over annotator votes rather than a single majority label, and making each prediction independently rather than relying on label frequency.

3.2 Prompt Design

Before settling on a production template, we ran a small-scale pilot comparing two candidate prompt formulations across a held-out sample of questions in each language, evaluated across several open-source models. Prompt variant 1 asked the model to return the correct option's full text with no explanation, while prompt variant 2 framed the task as a rule-based letter classifier, giving explicit if/then instructions ("if option (a) is correct, respond with: a") and requiring only a bare letter in response. Each model's zero-shot behaviour, with no instructional prompt at all, was also recorded as a reference point.

For the production system, a single prompt-building function is shared across training, validation, and test inference, so the fine-tuning distribution matches the evaluation distribution exactly. Each example is passed as a single user turn through the model's chat template, instructing it to act as a language-specific MCQ expert evaluating questions within corresponding national context. Naming the expected linguistic and cultural context explicitly was intended to reduce ambiguity for a single base model handling three typologically distinct languages within one fine-tuning run.

Finally, the shared template accommodates structural differences across the three subsets without splitting into separate per-language templates. A "Scenario" field is included only when a row provides one, a property specific to the Indonesian dataset, where dilemmas are grounded in real news-article contexts. For Sri Lankan rows, Option C and D are fixed Sinhala strings corresponding to "both answers are correct" and "neither answer is correct", reflecting the four-way answer space (A/B/Both/0) native to LKValues. These strings were kept in Sinhala rather than translated to a language-agnostic form, so the model is evaluated in the same linguistic register throughout the prompt.

```
You are a {Simplified Chinese | Indonesian | Sinhala} Expert
for answering multiple-choice questions.
You need to always evaluate these questions based on the
{Chinese | Indonesian | Sri Lankan} context and provide the
most accurate answer. You should only respond with the
correct option text, without any additional explanation or
commentary.
Scenario: {scenario}
Question: {question}
Option A: {option_a} <- only when the row has a non-empty
Scenario field (Indonesian)
Option B: {option_b}
Option C: {option_c}
Option D: {option_d}
Return only the correct option text. Expected output is
either the text of 'A', 'B', 'C', or 'D'.
```

Figure 1: Prompt

3.3 Model

3.3.1 Base Model Selection

Prior to finetuning, we benchmarked open-weight candidates under the Resource-Constrained Track's 8B parameter ceiling, zero-shot, on the full development set as shown in Table 2.

From this initial benchmark, we selected two candidates, Gemma-4-E4B-it and Qwen3.5-4B, for further evaluation. Both were assessed through official test-set submissions, and LoRA fine-tuning was then applied to each under comparable conditions, in order to compare how much headroom each model offered beyond its zero-shot baseline. Table 3 compares zero-shot and progressively fine-tuned test-set accuracy for both models across this comparison.

For Gemma-4-E4B-it, we tested fine-tuning with separate per-language adapters, as well as a Chinese-only fine-tuning configuration.

For Qwen3.5-4B, we first fine-tuned a LoRA adapter on Chinese data alone. Based on that result, we tried several further fine-tuning strategies on Qwen, including separate adapters per language, before continuing with additional stages incorporating Indonesian data and then augmented Sri Lankan data.

3.3.2 Final Submitted Model

Building on the base model selection above, our submitted system fine-tunes a single LoRA adapter on Qwen3.5-4B to jointly handle all three languages with one shared model, rather than training separate models per language.

Data: Training uses only the official labelled development sets, with no external data. For Indonesian, rather than collapsing the 5-annotator vote list to a single majority label, we train against the normalised vote distribution over A,B,C,D, so a 4-1 annotator split is treated differently from a 5-0 split. Chinese and Sri Lankan targets remain one-hot, for which this soft-label objective reduces to standard cross-entropy. The Sri Lankan task is cast as a 4-option MCQ by appending two fixed meta-options ("both correct"/ "neither correct", mapped back to the original answer space at prediction time.)

Curriculum training: Motivated by staged fine-tuning gains observed during base-model selection, we extend that idea into a formal six-stage curriculum, where each stage continues fine-tuning the adapter produced by the previous stage rather than restarting from the base model.

The design rationale follows directly from the base-model trials: staging one language at a time lets the adapter specialise before the next language is introduced, while two revisit stages (3 and 4) are included to counteract the forgetting that staged training risks, since gains from later stages could otherwise erode earlier-stage performance. The final joint stage then reconciles all three languages under one set of weights, with Chinese upweighted ($1.5\times$) to compensate for its lower absolute accuracy relative to the other two. The learning rate is lowered specifically for this final combined stage (5e-5 versus 1e-4 in all earlier stages).

Consistency-regularised objective: To encourage genuine reasoning over question content rather than reliance on answer position, each training example is paired at batch time with an option-permuted variant of itself; both variants receive full supervision against their (correspondingly permuted) target distribution. A Jensen-Shannon divergence term penalises disagreement between the model's predictions on the original and permuted versions after remapping to canonical letter order (JS is used over KL for its symmetry and boundedness).

For each example, the loss is

$$\mathcal{L}_{\text{lang}} = \frac{1}{2}\left(\mathcal{L}_{\text{original}} + \mathcal{L}_{\text{permuted}}\right) + \lambda \ \text{JS}\left(p_{\text{original}}, p_{\text{permuted}}\right), \tag{1}$$

where $\lambda = 0.5$ at every stage.

Training Details: LoRA rank 16, $\alpha = 32$, max sequence length 768, per-language batch size 8, AdamW, gradient clipping at norm 1.0, gradient checkpointing, bf16 precision. Loss is computed only at the single answer-token position. A 5-fold stratified cross-validation protocol selects the early-stopping iteration count (median best_iter across folds), after which a final model is refit on 100% of the data for that iteration count, avoiding permanent hold-out of data from an already small training pool.

Inference: Decoding is greedy (do_sample=False, max_new_tokens=40); any leading `<think>` span is stripped and the first valid standalone A/B/C/D token is extracted, with a fallback string match against option text. Unparseable outputs default to "A" to guarantee a prediction for every row, with the fallback rate logged per run.

3.4 Hardware and Software

All training and inference use Qwen3.5-4B fine-tuned with LoRA (Hu et al., 2022) via PEFT, implemented in PyTorch. Early base-model comparisons ran on Kaggle and Google Colab (T4 GPUs) and an Apple M1 Pro; the soft-label and consistency mechanisms, the curriculum experiments, and the final submitted run were trained on a local workstation with a single NVIDIA RTX 5090 (32GB VRAM), an Intel Core i9 (14th generation), and 64GB of DDR5 RAM.

4 Results

4.1 Base Model Comparison

All reported test-set numbers are official Codabench submission scores rather than local evalua-

tions. Table 2 reports zero-shot performance across all open-weight candidates benchmarked under the Resource-Constrained Track's 8B-parameter ceiling, evaluated on the full development set.

Qwen and Gemma variants achieved the strongest and most balanced macro-average accuracy across the three languages, while Mistral/Ministral variants frequently collapsed to single-letter answering and Llama3.1 underperformed specifically on Chinese and Sri Lankan.

Gemma-4-E4B-it achieved a higher zero-shot baseline than Qwen3.5-4B (macro 0.6912 vs. 0.6869), driven mainly by substantially higher Sri Lankan score (0.9235 vs. 0.8243). Its response to fine-tuning was mixed Figure 2. Per-language adapters scored below its own baseline (0.6635 vs. 0.6912), and Chinese-only fine-tuning left macro accuracy essentially unchanged (0.6904), with Chinese itself slightly lower (0.4937 vs. 0.5016). Adding an Indonesian stage (Chinese then Indonesian) did help, raising macro accuracy to 0.7040, Gemma's best result, with the gain coming from Indonesian (0.6485 to 0.6955). This remains below Qwen's first fine-tuned configuration (0.7531) and well below the submitted system (0.8062).

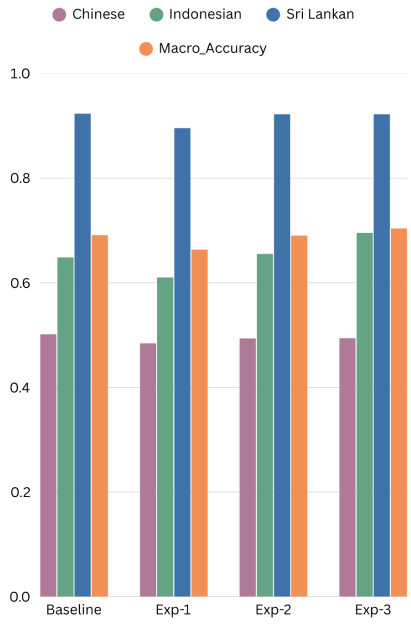


Figure 2: Progression of Gemma-4-E4B-it across fine-tuning configurations. Bars are ordered to match the corresponding rows in Table 3

Qwen3.5-4B, by contrast, improved substantially under matched fine-tuning. Its Chinese only fine-tune raised macro-accuracy from 0.6869 to 0.7531 (Chinese from 0.5694 to 0.7732), and successive curriculum stages added further gains, reaching 0.8044 macro accuracy.

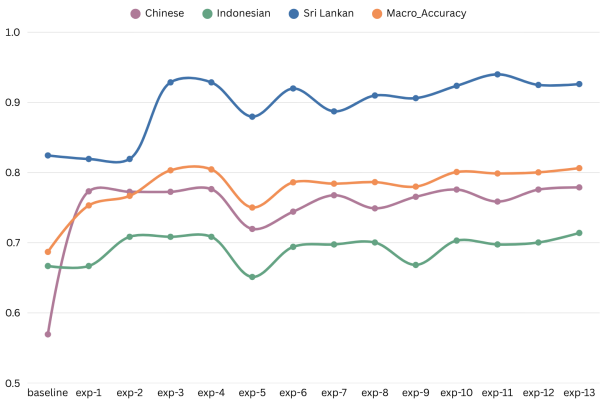


Figure 3: Progression of Qwen3.5-4B's macro-average accuracy from the initial fine-tuning comparison through curriculum training, in the same order as Table 3 and Table 5, ending with the final submitted system.

4.2 Curriculum Training Results

Table 4 shows the six-stage curriculum configuration used for the final submitted system.

Lowering the learning rate specifically for the final combined stage (5e-5 vs. 1e-4 in all earlier stages) raised test macro-accuracy from 0.8007 to 0.8062. This was the single change that improved all three languages simultaneously, and it was the configuration ultimately submitted.

4.3 Ablation Study

Table 5 presents the full ablation over training configurations tested, from simultaneous multilanguage training to the final curriculum design.

We attribute the final gain by comparing each step with its nearest predecessor in Table 5. Under simultaneous training, the soft-label and consistency objectives reach 0.7860–0.7863 against 0.7501 for a hard-label control (+3.6 points); this control ran on an earlier data pipeline that dropped the 72 Indonesian rows with tied votes, so the gap combines the objectives with that data change and we cannot separate them. Staging alone does not help: v1 (0.7799) is slightly below simultaneous training. The largest curriculum gain comes from adding the Indonesian revisit and the combined stage (v2, +2.1 points over v1, with Indonesian recovering from 0.6683 to 0.7030); because v2 changes both at once, we cannot attribute the gain to either alone. The combined-stage iteration count (v3) trades Chinese and Indonesian for Sri Lankan, and the Chinese loss weight (v4) stays

Model	Chinese	Indonesian	Sri Lankan	Macro Avg
google/gemma-3-4b-it	0.4215	0.6421	0.6305	0.5647
Qwen/Qwen2.5-7B-Instruct	0.4051	0.6175	0.5123	0.5116
unsloth/llama-3.1-8b-Instruct-bnb-4bit (8B)	0.2937	0.5710	0.3498	0.4048
mistralai/Mistral-7B-Instruct-v0.3	0.2975	0.4809	0.5271	0.4351
unsloth/Ministral-3-8B-Instruct-2512	0.2835	0.6120	0.2118	0.3691
unsloth/Ministral-3-3B-Instruct-2512	0.3582	0.4672	0.1970	0.3408
google/gemma-4-E4B-it	0.5114	**0.6967**	**0.8079**	**0.6720**
Qwen/Qwen3.5-4B	**0.5342**	0.6448	0.6789	0.6193
MERaLiON-LLaMA-3-8B-Instruct	0.3171	0.6376	0.6424	0.5323

Table 2: Zero-shot macro-accuracy of open-weight candidates under the Track-1 8B ceiling, evaluated on the full 1,359-row development set.

Model	Variant	Chinese	Indonesian	Sri Lankan	Macro_Accuracy
google/gemma-4-E4B-it	baseline	0.5016	0.6485	0.9235	0.6912
	lora fine-tuned (separate adapters per language)	0.4844	0.6104	0.8959	0.6635
	lora fine-tuned (chinese only)	0.4937	0.6553	0.9222	0.6904
	lora fine-tuned (chinese then indonesian)	0.4943	0.6955	0.9222	0.7040
Qwen/Qwen3.5-4B	baseline	0.5694	0.6668	0.8243	0.6869
	lora fine-tuned (chinese only)	0.7732	0.6668	0.8193	0.7531
	lora fine-tuned (chinese then indonesian)	0.7725	0.7084	0.8193	0.7667
	lora fine-tuned (chinese then indonesian then aug sri lankan)	0.7725	0.7084	0.9284	0.8031
	lora fine-tuned (chinese then indonesian then aug sri lankan then aug chinese)	0.7763	0.7084	0.9284	**0.8044**

Table 3: Codabench test-set accuracy for the two candidate base models across successive fine-tuning variants. Contrast with Table 2, which reports development-set zero-shot accuracy.

within 0.0005 of v2. Lowering the combined-stage learning rate (v5) adds roughly half a point (0.8062 vs. 0.8007 for v2 and 0.8002 for v4) and is the only change that improves all three languages relative to both. v6 shows this gain is specific to the final stage: applying 5e-5 to every stage, together with per-stage iteration selection, lowers macro accuracy to 0.7911. The final configuration (v5) reaches 0.7788 / 0.7139 / 0.9260 on Chinese / Indonesian / Sri Lankan (macro 0.8062).

5 Discussion

A stronger starting point is not a stronger substrate: Gemma-4-E4B-it held the better zero-shot macro (0.6912 vs. 0.6869), driven almost entirely by Sri Lankan (0.9235 vs. 0.8243). However, the ordering reversed under matched LoRA conditions: Qwen's Chinese-only fine-tune raised macro accuracy from 0.6869 to 0.7531, while the matched Gemma configuration left it unchanged (0.6912 to 0.6904), and Gemma's per-language-adapter variant fell below its own zero-shot score (0.6635 vs. 0.6912). Gemma's best configuration, Chinese then Indonesian (0.7040), did improve on its baseline, but it remained below every fine-tuned Qwen configuration. The cause was not isolated, though the pattern is consistent with Gemma's Sri Lankan advantage arising from instruction-tuning that already fits this item format closely, leaving little headroom and making the adapter more likely to disturb an existing solution than to improve it. The practical lesson for constrained tracks is that zero-shot benchmarking is a poor proxy for adaptability, and that base-model choice should be made on measured fine-tuning response rather than on baseline rank.

The development set does not predict the test set: Zero-shot Sri Lankan accuracy for Gemma-4-E4B-it is 0.8079 on the development set (Table 2) but 0.9235 on the test set (Table 3); the Gemma-Qwen ordering on Indonesian reverses (0.6967 vs. 0.6448 on development; 0.6485 vs. 0.6668 on test). With 203 labelled Sri Lankan rows, a development estimate carries a standard error about 2.8 points, and the two splits are not guaranteed to be matched in difficulty. This gap was visible in the development logs and directly shaped the development stages: we adopted cross-validated iteration selection with a full-data refit rather than a fixed held-out split, and treated development scores as a filter for candidates rather than as a predictor of final ranking.

Why Sri Lankan is the easiest subset: Sri Lankan accuracy (0.9260) exceeds Chinese (0.7788) by 15 points despite having the fewest training rows and the lowest-resource language. Part of this gap reflects label priors rather than difficulty: on the development sets, always predict-

Stage	Language	Iterations	Learning rate	Loss weight	Notes
0	zh	25	1e-4	1.0	fresh LoRA from base model
1	id	25	1e-4	1.0	continues stage 0
2	si	25	1e-4	1.0	continues stage 1
3	zh	25	1e-4	1.0	Chinese revisited
4	id	25	1e-4	1.0	Indonesian revisited (prevents forgetting)
5 (combined)	zh+id+si	25	5e-5	zh 1.5, id 1.0, si 1.0	simultaneous; final submitted checkpoint

Table 4: Training stages and hyperparameters.

Configuration	Chinese	Indonesian	Sri Lankan	Macro_Accuracy
simultaneous, hard-label (no soft-label/ consistency)	0.7196	0.6512	0.8795	0.7501
simultaneous + soft-label + consistency (λ=0.5, CV iters)	0.7442	0.6941	0.9197	0.7860
simultaneous + soft-label + consistency (λ=0.6, refit 50)	0.7676	0.6975	0.8871	0.7841
simultaneous + soft-label + consistency (λ=0.6, CV iters)	0.7489	0.7003	0.9097	0.7863
curriculum v1 (zh, id, si, zh; no combined stage)	0.7654	0.6683	0.9059	0.7799
curriculum v2 (+id revisit + combined stage, zh-w 1.5)	0.7757	0.7030	0.9235	0.8007
curriculum v3 (combined stage-iters 25→40)	0.7586	0.6975	0.9398	0.7986
curriculum v4 (zh-weight 1.5→1.3)	0.7757	0.7003	0.9247	0.8002
curriculum v5 (combined stage lr 5e-5)	**0.7788**	**0.7139**	**0.9260**	**0.8062**
curriculum v6 (per-stage CV + lr 5e-5 everywhere)	0.7710	0.7003	0.9021	0.7911

Table 5: Ablation over training configurations. The final submitted configuration (curriculum v5) is shown in bold.

ing the most frequent label gives 52.2% for Sri Lankan (A, 106 of 203) but 37.2% for Chinese (D, 294 of 790), so our margin over that reference is nearly identical (+40.4 vs. +40.7 points). We also suspect, but did notv erify, that Sri Lankan items are easier to discriminate because they contrast two statements, whereas every Chinese option is constructed to be defensible. High zero-shot Sri Lankan accuracy (0.8243 for Qwen3.5-4B, 0.9235 for Gemma-4-E4B-it) is consistent with this but does not establish it. Per-country scores should therefore not be read as a ranking of how well the model "understands" each culture.

A tension in the task, and in our own system: The benchmark is motivated by the claim that aggregating preferences into an average annotator marginalises populations, yet Indonesian systems are scored against the annotator consensus. A system optimised for that metric is being rewarded for reproducing a majority, the behaviour the framing objects to, relocated from a global to a national scale. Our soft-label objective partially resists this by training on the full vote distribution, so that a 4-1 split is not treated as a 5-0 split, but the evaluation still collapses to consensus at scoring time. Retaining vote distributions in the released data, and reporting a distributional metric alongside consensus accuracy, would let future editions of this task measure pluralism directly rather than by proxy.

6 Conclusion

We presented our submission to Track 1 of the PlurVA-LLM 2026 shared task, unifying three culturally and structurally distinct value-alignment datasets Chinese, Indonesian, and Sri Lankan into a single system built on Qwen3.5-4B, combining label-space unification, staged curriculum fine-tuning, and an option-permutation consistency objective to reduce answer-position bias. Although Gemma-4-E4B-it achieved stronger zero-shot and baseline accuracy, Qwen3.5-4B responded far more consistently to LoRA fine-tuning, and our experiments showed that sequential, language-by-language curriculum training outperformed joint multi-language fine-tuning from initialisation, with a reduced learning rate in the final joint stage producing balanced gains across all three languages. Our final submitted system reached a macro-average accuracy of 0.8062 on the hidden test set, reflecting improvement distributed across Chinese, Indonesian, and Sri Lankan rather than concentrated in a single language, consistent with the shared task's emphasis on cross-country balance. Future work could explore larger-scale data augmentation, alternative consistency objectives, and extending this curriculum approach to additional low-resource cultural contexts.

Limitations

While the proposed curriculum fine-tuning approach demonstrates strong accuracy gains, it has several limitations.

Only 1359 labelled rows in total (790 Chinese, 366 Indonesian, 203 Sri Lankan) were used for fine-tuning, with no external data. This is a small pool for adapting a 4B-parameter model, and the Sri Lankan subset in particular (203 rows) is very small, a risk our experiments logs themselves flag.

Beyond data scarcity, our evaluation reports only aggregate macro accuracy per language and does not yet localise where the remaining errors occur by country, by answer type, or by strength of annotator disagreement for Indonesian, an analysis that would be needed to target further improvements precisely. We also did not test whether naming the target value in the prompt improves accuracy.

In addition, the 8B parameter cap imposed by the Resource-Constrained Track ruled out potentially stronger, larger models. The results reported here are specific to what is achievable under that constraint and may not reflect the ceiling of value-alignment performance more broadly.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Alejandro Benito-Santos and Adrian Ghajari. 2025. Beyond averages: Learning with annotator disagreement in STS. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35551–35557.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *Proceedings of the First Conference on Language Modeling*.

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.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2591–2597.

Blessed Guda, Lawrence Francis, Gabrial Zencha Ashungafac, Carlee Joe-Wong, and Moise Busogi. 2025. Quantifying and mitigating selection bias in LLMs: A transferable LoRA fine-tuning and efficient majority voting approach. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025*.

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.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6997–7013.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

J. A. T. K. Jayakody, T. S. K. Gamlath, W. A. N. Lasantha, K. M. K. P. Premachandra, Anupiya Nugaliyadde, and Yashas Mallawarachchi. 2016. Mahoshadha, the sinhala tagged corpus based question answering system. In *Advances in ICT for Emerging Regions*. Springer.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for hindi LLMs. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. 2025. Challenges in adapting multilingual LLMs to low-resource languages using LoRA PEFT tuning. In *Proceedings*

of the First Workshop on Challenges in Processing South Asian Languages, pages 217–222.

Y. M. W. H. M. R. P. J. R. B. Kiridana and Gihan Dias. 2025. Developing a question answering system for the sri lankan school education system. In *Proceedings of Applied Data Science and Artificial Intelligence Symposium 2025*, pages 55–56.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew M. Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37.

Olga Loginova, Oleksandr Bezrukov, Ravi Shekhar, and Alexey Kravets. 2025. Addressing blind guessing: Calibration of selection bias in multiple-choice question answering by video language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 3216–3246.

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.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Aleksandr Perevalov, Andreas Both, and Axel-Cyrille Ngonga Ngomo. 2024. Multilingual question answering systems for knowledge graphs – a survey. *Semantic Web*, 15:1–36.

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.

Janani Ranasinghe and Ruvan Weerasinghe. 2025. Question answering in a low-resource language: Dataset and deep learning adaptations for sinhala. In *Deep Learning Theory and Applications*, pages 336–352. Springer Nature Switzerland.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, et al. 2024. INCLUDE: Evaluating multilingual language understanding with regional knowledge. *Preprint*, arXiv:2411.19799.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*.

Ixak Sarasua, Ander Corral, and Xabier Saralegi. 2025. DIPLomA: Efficient adaptation of instructed LLMs to low-resource languages via post-training delta merging. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24898–24912.

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. In *Proceedings of the 41st International Conference on Machine Learning*.

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.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. Leveraging annotator disagreement for text classification. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10, Trento. Association for Computational Linguistics.

Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, Jie Zhou, and Yue Zhang. 2023. Consistency regularization training for compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1294–1308.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

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*, pages 19834–19852.