South Asian Voices in LLMs: Culturally Aware Multilingual Instruction Fine-Tuning for South Asian Low-Resource Languages

Anonymous ACL submission

Abstract

Can large language models (LLMs) truly understand and represent the regional-wise rich cultural and linguistic diversity? Addressing this critical question, our study aims to develop a culturally adaptive multilingual instruction dataset to fine-tune LLMs in order to enhance their cultural alignment, multilingual fluency, and instruction-following capabilities across 15 South Asian low-resource languages (LRLs). We systematically constructed the South Asian Instruction Dataset (SAID) by combining automated LLM-based semantic categorization, human-in-the-loop cultural tagging, and country-specific localization using state-of-the-art multilingual LLMs. This dataset spans eight SAARC countries and covers ten culturally relevant domains. We employed parameter-efficient LoRA fine-tuning on the LLaMA 3.1 Instruct model and conducted a comprehensive evaluation combining automated LLM judgment with human expert assessment. The resulting fine-tuned model, which we call SAID-LLaMA 3.1 Instruct, demonstrates substantial improvements over the base LLaMA 3.1 Instruct model in generating culturally aligned, factually accurate, and linguistically fluent responses for high- and mid-resource South Asian languages. Theoretically, this work advances understanding of how cultural adaptation and multilingual fine-tuning can enhance LLM performance in low-resource contexts. Practically, it provides a high-quality, culturally grounded instruction dataset, fine-tuning and evaluation methodology that can guide the development of more inclusive AI systems for South Asia.

1 Introduction

South Asia is one of the most linguistically and culturally diverse regions in the world, home to hundreds of languages spanning multiple language families, dialects, and scripts. This linguistic tapestry is deeply intertwined with a rich mosaic of cultural traditions, histories, and social practices, making the region a uniquely complex and vibrant context for natural language processing (NLP) (Ramesh et al., 2022; Guzmán et al., 2019; Kunchukuttan et al., 2020). Despite advances in NLP, most South Asian languages remain underrepresented in research and technology. As LRLs, they lack annotated datasets, pretrained models, and culturally relevant resources. This limits the inclusivity of language technologies and makes it difficult for AI systems to reflect South Asia's cultural diversity.

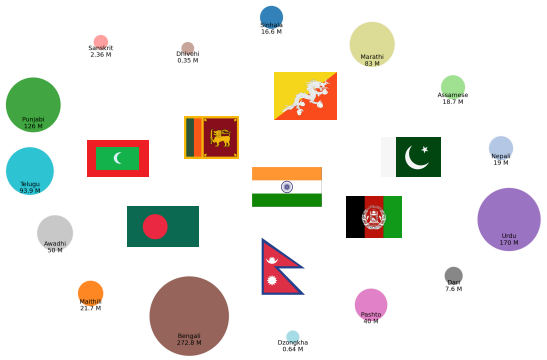

Figure 1: Bubble diagram of 15 South Asian languages, where each circle's size is proportional to its number of speakers (in millions), and flags of the SAARC countries (Bhutan, Sri Lanka, Maldives, Pakistan, India, Afghanistan, Bangladesh, Nepal).

But what exactly do we mean by **"Culture"?** It is difficult to define culture precisely, as it is not a static concept but a continuously evolving and dynamic entity that shapes the entire way of living for a particular group of people. In this work, we define culture as the collective expressions, knowledge, practices, artifacts, local languages, and historical experiences that shape the identity of a particular country or community. To operationalize this broad and complex concept within our research, we decomposed the culture of each South Asian country

(SAARC countries)[1] into ten thematic categories, or cultural labels. These labels include literature, entertainment, language and grammar, history and religion, people, geography, food and beverages, flora and fauna, sports, and festivals.

The motivation behind this research stems from the pressing need to bridge this gap by developing culturally rich, multilingual instruction datasets and fine-tuning large language models (LLMs) that truly embody the linguistic and cultural diversity of South Asia. Existing datasets focus on high-resource languages (HRLs) or use machine translation, missing cultural specificity. Current evaluation metrics also overlook local fluency and cultural accuracy, which are crucial for AI in diverse contexts (Rystrøm et al., 2025; AlKhamissi et al., 2024).

To address these challenges, our study asks: *How can we systematically construct a multilingual, culturally adaptive instruction dataset that covers a broad spectrum of South Asian LRLs? And, to what extent can fine-tuning LLMs on such datasets to improve their cultural alignment, multilingual fluency, and instruction-following capabilities? Moreover, how can we rigorously evaluate these models to ensure they truly reflect the cultural and linguistic diversity inherent to South Asia?*

Our methodology unfolds in several key stages. We begin with extensive categorization and semantic labeling of Stanford-Alpaca dataset [2] (Taori et al., 2023) using a combination of automated LLM-based classification and manual human annotation to separate language-related, culturally relevant, and general instructions. This is followed by thematic cultural tagging across multiple South Asian cultural domains, which is then supported by keyword-driven few-shot prompting and human validation to ensure accuracy. We then localize this dataset to 8 SAARC countries using state-of-the-art (SOTA) multilingual LLMs, Command R and Llama 4 Maverick, incorporating culturally specific prompts to tailor outputs relevant to each nation's unique context (Singh et al., 2024a; Wu et al., 2023).

To prepare the data for multilingual fine-tuning, we distill a high-quality subset (Zhou et al., 2023) of culturally rich instructions which are equally distributed among the 8 countries and translated it into 15 South Asian LRLs, creating the comprehensive South Asian Instruction Dataset which we call **the SAID**. The dataset's cultural and linguistic integrity was rigorously maintained through a multi-layered human-in-the-loop quality control process. Finally, we employed parameter-efficient fine-tuning with LoRA adapters on the LLaMA 3.1 Instruct model,[3] enabling efficient adaptation while preserving model capacity (Dettmers et al., 2023; Li et al., 2023).

Our evaluation framework combined automated LLM-based judgment (Kim et al., 2024) with large-scale human expert evaluation across multiple languages and cultural contexts. We designed a mini gold-standard test set with detailed rubrics focusing on instruction-following quality, factuality and cultural alignment, and multilingual fluency.

The major findings reveal that fine-tuning on the SAID dataset yields models with moderate to strong improvements in generating culturally aligned, factually accurate, and linguistically fluent responses, particularly for high- and mid-resource South Asian languages.

Beyond our localized test set, these gains extend to standard downstream tasks. On broader multilingual benchmarks for translation, summarization and question answering, the SAID-fine-tuned model showed measurable improvements across key automatic metrics and factual accuracy, with notable progress in most LRLs tested. The human evaluations consistently rated the SAID LLaMA 3.1 Instruct more favorably than LLM-based-judgments, underscoring the indispensable role of native expert judgment in culturally nuanced LLM evaluation.

Our contributions are as follows:

1. We curate and release the first culturally adaptive multilingual instruction dataset spanning the 8 SAARC countries in South Asia, 15 South Asian LRLs, covering 10 diverse cultural domains (see Figure 1).

2. We demonstrate a novel pipeline combining LLM-based semantic categorization, cultural tagging, and country-specific localization with rigorous human-in-the-loop quality control.

3. We show the effectiveness of parameter-efficient LoRA fine-tuning on LLaMA 3.1 using culturally rich multilingual data to enhance model cultural alignment and multilingual fluency.

[1] SAARC
[2] Stanford Alpaca
[3] Base Model-Meta LLamA 3.1 Instruct

4. We establish a comprehensive, dual-mode evaluation framework integrating automated LLM judgments and native human expertise with detailed cultural rubrics.

Our main contribution is a systematic, reusable pipeline for building country-specific, culturally inclusive instruction-tuning datasets. By focusing on data creation and evaluation methods rather than new algorithms, this work lays the groundwork for more equitable and culturally aware multilingual AI in South Asia and beyond.

2 Related Work

Multilingual Instruction Datasets for LRLs. Instruction-tuning in LRLs has been boosted by several multilingual datasets. MURI (Köksal et al., 2024) introduces reverse instruction and translation to generate pairs from existing texts. Aya (Singh et al., 2024b) aggregates 513M instances across 114 languages. TaCo (Upadhayay and Behzadan, 2023) uses curriculum learning with translated instructions. BayLing 2 (Zhang et al., 2024b) and LinguaLIFT (Zhang et al., 2024a) enable cross-lingual transfer with alignment layers and code-switching.

LLMs as Zero-Shot or Few-Shot Text Classifiers. LLMs can act as powerful text classifiers in low-resource settings. Wang et al. (2023) evaluate GPTs as zero-shot classifiers. In healthcare,Guo et al. (2024) find LLMs outperform SVMs and transformers. Patwa et al. (2024) use few-shot learning and synthetic data to improve classification. Vajjala and Shimangaud (2025) and Parikh et al. (2023) examine prompt strategies and adaptation across domains and languages.

South Asian NLP. Resources for South Asian languages have expanded with corpora like Samanantar (Ramesh et al., 2022), FLORES (Guzmán et al., 2019), IndicNLP (Kunchukuttan et al., 2020) and instruction dataset collections including IndicInstruct (Gala et al., 2024). Bactrian-X (Li et al., 2023) and Aya (Singh et al., 2024b) use translation crowdsourcing, human and GPT-based annotation which includes some South Asian languages. The SinLlama project (Aravinda et al., 2025) released the first open-source Sinhala LLM to address the longstanding gap for Sinhala. All these work highlight the growing diligence and progress of the South Asian NLP community in bridging the linguistic divide in large language models.

Dataset Localization Using Large Language Models. LLMs themselves can localize data. LAMINI (Wu et al., 2023) uses GPT-4 and few-shot prompting for multilingual instruction generation. Singh et al. (2024a) explore culturally sensitive paraphrasing using GPT-3.5. BLOOM (Le Scao et al., 2023) and XLM-R (Conneau et al., 2020) support such efforts with multilingual pre-trained architectures.

Cultural Adaptation and Multilingual Fine-Tuning of LLMs. Cultural alignment enhances model relevance. Xu et al. (2024) propose CultureSPA with cultural prompts. Hadar-Shoval et al. (2024) and Rystrøm et al. (2025) highlight value divergence. Anthropological prompting (AlKhamissi et al., 2024) and culture-sensitive rewriting (Singh et al., 2024a) illustrate alignment challenges. AmbigNLG (Niwa and Iso, 2024) shows the utility of human-in-the-loop workflows.

Multilingual Fine-tuning with PEFT or Small Datasets. Li et al. (2023); Dettmers et al. (2023) enables high performance with compact data. LegalQA-bloom-560m,[4] and LIMA (Zhou et al., 2023) show that small, curated datasets can rival large, noisy ones. Lima-X (Weber et al., 2024), Guanaco (Dettmers et al., 2023), and Flacuna (Ghosal et al., 2023) extend this across languages with <50k dialogues.

Prior work confirms that (i) multilingual instruction corpora enhance low-resource performance when culturally relevant, (ii) LLMs themselves can act as classifiers and bootstrap annotation but need human verification, and (iii) compact, high-quality datasets may outperform massive noisy ones, especially with PEFT techniques. However, no existing resource simultaneously targets the full SAARC regional spectrum and encodes explicitly South-Asian cultural knowledge. Our study fills this gap by releasing the first 15 low-resource language, culturally adaptive instruction set for South Asia and by testing PEFT LoRA on LLaMA-3.1 Instruct across 10 cultural domains, thereby extending the current insights to a new linguistic-cultural dimension.

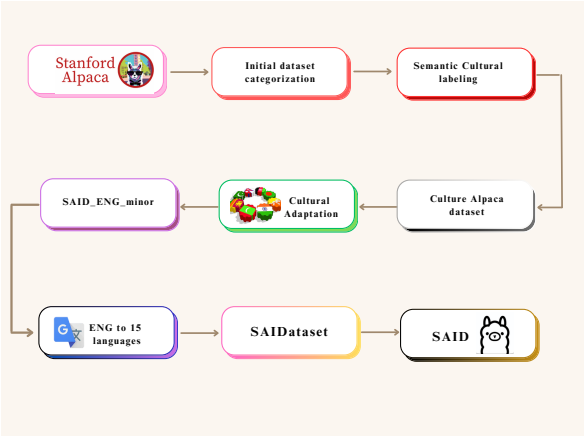

Figure 2: Overview of the SAID dataset and process for SAID-LLaMA model creation.

3 Dataset Curation and Localization

This section describes the systematic approach taken to construct a culturally adaptive, multilingual instruction dataset tailored for low-resource South Asian languages. Please see Figure 2.

3.1 Initial Dataset Categorization

We began with the Alpaca cleaned dataset(Taori et al., 2023), a comprehensive English instruction dataset containing 52,000 instances. To structure this data for cultural adaptation, we initially leveraged the zero-shot classification capabilities of the DeepSeek R1 model, which categorizes instructions into 3 high-level groups: (1) **Language-related**: Instructions concerning grammar, translation, and text manipulation. (2) **Culture-relevant**: Instructions related to literature, places, people, plants and animals, sports, festivals and traditions, history and religion, entertainment, food and beverages, and geography. (3) **General**: Instructions that do not fit into the previous two categories. The classification prompt instructed the LLM to decide based on the content whether an instruction belonged to language, culture, or general categories. This initial separation resulted in 17,400 language-related, 5,300 culture-relevant, and 28,800 general instructions. (See Appendix A)

3.2 Semantic Cultural Tagging

To refine the cultural dimension, we further classified the language-related and culture-relevant instructions into ten thematic categories: literature, entertainment, language and grammar, history and religion, people, geography, food & beverages, flora & fauna, sports, and festivals. A category named “other” was included for uncategorizable instances. We compiled keyword lists for each category and used few-shot prompting with DeepSeek R1 to assign cultural tags.

Domain-specific tagging and annotation of content with elements is inherently challenging. Often an instruction can belong to multiple categories or lie on the boundary of several cultural elements. Human-in-the-loop processes are essential to handle these ambiguities (Niwa and Iso, 2024). Following automatic tagging, we performed manual validation and adjustment on the literature and other categories. This involved reassigning instructions where the LLM misclassified or when cultural localization was questionable (Niwa and Iso, 2024), especially for topics like machine learning books, which were retained in the "other" category due to their non-localizable nature.

3.3 Human Annotation and Dataset Finalization

To finalize the English instruction dataset suitable for South Asian cultural localization, we engaged two South Asian domain expert annotators who are university students. Their responsibilities included verifying category assignments and correcting any mislabels and assisting us with the evaluation scoring. During this process, considerable number of instances from "literature" and "other" categories that shows lesser localization potential are reassigned to the "general" category. This yields the `culture_alpaca_dataset` with 7,833 curated instructions. (See Appendix A)

3.4 Cultural Adaptation to South Asian Countries

The core innovation of our approach is the cultural adaptation of the `culture_alpaca_dataset` to eight South Asian countries forming the SAARC region: Sri Lanka, India, Pakistan, Nepal, Bhutan, Bangladesh, Afghanistan, and Maldives. We crafted culturally-aware country-specific few-shot prompts instructing the LLMs to localize content retaining English output but incorporating culturally relevant references, terminology, and examples unique to each country (Singh et al., 2024a). This reduces the burden on human experts by letting the LLMs do a first pass localization.

A cruicial consideration in dataset localization is the choice of LLM. To accurately produce localized text, the LLM must have strong abilities in the

[4]LegalQA

target country or culture. For this task, we utilized two SOTA multilingual LLMs, Command R and Llama 4 Maverick, selected for their robustness in low-resource settings, multilingual capabilities, and various creativity levels.

The localized datasets were independently reviewed by eight native speakers from SAARC countries, who are undergraduate and graduate students, fluent in their local languages, the national language, and English. Reviewers fact-checked the localized data and assessed based on cultural accuracy and fluency, providing quality ratings between 7 and 10 out of 10 per country. Because some instructions were challenging to localize, the final localized datasets for each country varied in size. These were merged into a comprehensive SAID-English dataset comprising 53,760 instructions after going through a critical deduplication process to ensure that there are no instances with similar instructions or response ideas.

3.5 Creation of a High-Quality Multi-cultural Subset

To enhance dataset quality for multilingual training, we distilled a smaller, high-quality subset of 1,510 instances named `SAID_ENG_minor` following the rigorous LIMA methodology (Zhou et al., 2023). This approach emphasizes that a carefully curated, smaller dataset selected for clarity, informativeness, politeness, and controlled length (1,200–4,096 characters) can yield better results which we believe is ideal for low resource setting. The un-selected instructions were reserved for making the test set and future rounds of model development, as noted in our Limitations section. See Table 1 for dataset statistics.

To enrich cultural relevance, we supplemented this subset with 100 newly created instructions based on real South Asian user conversations sourced from Quora [5]. These were selected to increase coverage of country-specific, nuanced cultural phenomena and user concerns that are often missing from conventional datasets. Quora data accounted for a small portion of the subset and was incorporated to bring a real-world, user-centric dimension, yet it does not meaningfully confound our experimental outcomes.

[5] Quora.com

3.6 Translation and Multilingual Dataset Assembly

The `SAID_ENG_minor` dataset, structured with instruction, input, and output pairs consistent with the Alpaca format, was machine translated into 15 South Asian LRLs including Sinhala, Nepali, Maithili, Punjabi, Assamese, Sanskrit, Urdu, Bengali, Dhivehi, Pashto, Dari, Awadhi, Marathi, Telugu, and Dzongkha using the Google Translate API. Along with the original English data, this resulted in a comprehensive multilingual dataset spanning 16 languages, collectively forming the SAID-Multilingual dataset with over 24,000 instances. This dataset serves as the foundation for our multilingual instruction fine-tuning.

3.7 Quality Control

Ensuring the quality, cultural authenticity, and overall integrity of the SAID was central to our methodology. We implemented a multi-layered human-in-the-loop quality control process integrated across all key phases of dataset development, including initial categorization, semantic cultural tagging, localization, and multilingual translation.

During initial dataset categorization and semantic cultural label tagging, we involved two South Asian domain experts to ensure category assignments were accurate and culturally appropriate for South Asian contexts. To promote consistency, we provided detailed annotation guidelines and conducted consensus discussions for ambiguous or conflicting cases. Difficult borderline or ambiguous cases were flagged for further review or exclusion, and categories such as "science" and "others" were explicitly removed from the culturally-adaptive subset. With the help of the 2 experts, we finalized the **Culture_Alpaca_dataset** of 7,833 instances.

After localization to each of the eight SAARC countries using SOTA multilingual LLMs, we engaged a broader team of 8 native-speaking reviewers: each fluent in one or more of the target languages to conduct an independent review of the localized data. They evaluated each country's data for cultural accuracy, linguistic naturalness, and appropriateness. Quality ratings of 7–10 out of 10 guided iterative improvements, ensuring high standards across all countries.

The translation quality of the LRLs was carefully monitored during the machine translation of the `SAID_ENG_minor` dataset into 15 South Asian

Country	Localized	SAID_ENG-minor
Afghanistan	5550	187
Bangladesh	7568	98
Bhutan	6183	220
India	7564	203
Maldives	7491	202
Nepal	6008	200
Pakistan	5835	200
Sri Lanka	7556	200

Table 1: Distribution of localized instances and subset counts in `SAID_ENG_minor` across South Asian countries.

languages using the Google Translate API. Here we again involved native speakers for quality assurance. Only minor corrections were needed, underscoring the high accuracy of the machine translation process for the target languages. Datasets were only finalized after these human validation steps.

4 Evaluation and Analytical Framework

Evaluating LLMs fine-tuned on culturally grounded data requires two complementary perspectives such as how well the model follows localized instructions, and whether those gains can transfer to standard multilingual tasks.

4.1 Dual-Mode, Culture-Focused Evaluation

Our evaluation proceeded in 2 modes: (1) **LLM-as-a-Judge** using Prometheus-7b-v2.0 (Kim et al., 2024). The Prometheus model scored responses on a 0–5 scale against detailed rubrics crafted for each question, ensuring rigor and standardization in automated evaluation. (2) **Human Expert Evaluation**: We engaged eight native speakers, each fluent in at least two of the 15 South Asian languages under study, to perform independent, blind evaluations of the fine-tuned and base model outputs. The human evaluators used the exact same rubric and scoring criteria as Prometheus to maximize consistency and comparability between human and automated assessments.

To test the model effectively, we compiled a **mini gold standard dataset** of 150+ open-ended question-answer pairs, carefully selected from trusted sources such as the Aya Evaluation Suite (Singh et al., 2024b), L3Cube-IndicNLP datasets (Deode et al., 2023) and the instruction data which wasn't used to create the `SAID_ENG_minor` from the **Culture_Alpaca_dataset**. Each prompt was paired with a reference answer and a corresponding rubric exemplifying three core evaluation criteria: (1) Instruction-Following Quality, (2) Factuality and Cultural Alignment, and (3) Multilingual Fluency. This rubric-driven approach allowed both human annotators and Prometheus to score responses systematically, promoting a detailed and multidimensional evaluation of model quality.

4.2 General-Purpose Benchmark Evaluation

To verify that SAID fine-tuning does not trade off general capability and can even enhance it, we therefore benchmark the base and fine-tuned models on three widely-used suites that cover translation using FLORES-200 (NLLB Team, 2022), summarization using Aya Evaluation Suite (Singh et al., 2024b) and question-answering using L3Cube-IndicNLP QA (Deode et al., 2023).

4.3 Statistical Metrics for Quantitative Insights

For both human and LLM evaluations, we computed: (1) **Mean Score**: reflects the average quality rating across all responses and provides a central tendency measure of model performance; (2) **Standard Deviation (Std Dev)**: measures the variability in scores to indicate consistency or fluctuation in model output quality. (3) **Percentage of Examples Rated** ≥ 3: denotes the proportion of responses considered acceptable or better. (See Appendix A).

4.4 Semantic Categorization and Cultural Tagging Consistency

Beyond evaluating raw output quality, it was critical to assess the LLM's ability to correctly interpret and categorize instructions with cultural relevance. We examined semantic categorization consistency to verify whether the model aligns with human judgment in distinguishing general, language-related, and culture-specific content. Further, we analyzed the model's proficiency in tagging cultural labels. This evaluation aimed to gauge the model's capacity to act as a classifier.

4.5 Comparative Localization Analysis Between the SOTA LLMs

To benchmark cultural adaptation capabilities more comprehensively, we compared two leading LLMs, Command R and Llama 4 Maverick, on their ability to localize a substantial dataset for two culturally and linguistically distinct South Asian countries: India and Sri Lanka. This comparison was motivated by the need to evaluate how models handle localization in both resource-rich and resource-

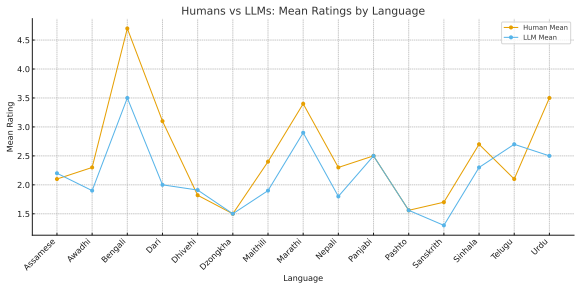

Figure 3: Heatmap of Mean Ratings by Language (Human vs LLM).

limited environments, reflecting real-world diversity in cultural and NLP ecosystem maturity.

4.6 Cross-Regional Cultural Transferability

To examine the finetuned model's capacity to transfer South Asian cultural knowledge beyond its core domain, (Weber et al., 2024) we used 30 open-ended questions from the **Culture_Alpaca_dataset** and machine translated into 6 non-South Asian languages: three high-resource (French, Arabic, Chinese) and three low-resource (Uyghur, Hawaiian, Zulu). Please see Appendix A).

5 Experiments

Our comprehensive evaluation of the culturally fine-tuned SAID LLaMa 3.1 Instruct model across multiple South Asian languages and diverse evaluation contexts yielded several critical insights and challenges inherent in multilingual, culturally nuanced NLP systems. As shown in Figure 3, the chart illustrates the mean rating comparisons between human evaluators and the LLM across various South Asian languages.

5.1 Model Fine-tuning with LoRA Adapters

We fine-tuned the LLaMA 3.1 Instruct model using parameter-efficient LoRA adapters (Li et al., 2023). LoRA modules were inserted into all attention layers, so only a small fraction of model parameters were updated during training while the rest of the model remained frozen. We used 4-bit QLoRA quantization with the AdamW optimizer and a global batch size of 256, training for 3 epochs on 4,096-token sequences. All training was run in bfloat16 mixed precision on a single NVIDIA A100 80GB GPU and completed in approximately two hours.

5.2 Aggregate Performance Summary

The fine-tuned model demonstrated moderate overall quality in generating culturally relevant, factually accurate, and linguistically fluent responses. Human evaluators rated the outputs with an average mean score of 2.58 (±1.23) on a 0–5 scale, with approximately 49% of responses meeting or exceeding an acceptable quality threshold (score ≥ 3). This indicates a reasonable degree of success in instruction adherence, factual correctness, and cultural alignment. In comparison, the automated evaluation by Prometheus-Eval 7B exhibited a more conservative stance, assigning a lower mean score of 1.86 (±0.97) and only about 24% of responses meeting the acceptability benchmark. This discrepancy highlights the complexities automated judges face when assessing nuanced cultural and linguistic aspects, and underscores the indispensable role of human expert judgment for culturally sensitive evaluation. (See Appendix A).

5.3 General-Purpose Task Results

The fine-tuned model showed improvements on the FLORES-200 multilingual translation benchmark, with a notable increase in BLEU and a reduction in WER compared to the base model. The best performance was observed in Maithili, which saw substantial improvements in BLEU(0.02-0.16), NIST(0.27-2.03), chrF (0.29-0.51). For Aya Evaluation Suite summarization task, the fine-tuned model showed improvements in key metrics: METEOR (16.19%), BLEU (11.59%), and Cosine Similarity (8.46%). Hindi demonstrated the highest gains in BLEU and METEOR. In the L3Cube-IndicNLP question-answering task, the fine-tuned model demonstrated improvement in factual accuracy which rose by 36.36% and a significant increase in response quality, reinforcing the model's enhanced performance across diverse tasks. These extended evaluations underscore the broad value of the SAID-fine-tuned model, reinforcing that the dataset's culturally inclusive and multi-domain nature leads to substantial improvements across various tasks. (See Appendix A)

5.4 Language-Specific Variability

We separated the 15 chosen South Asian languages into three groups as high-resource, mid-resource, low-resource within the South Asian domain. Let us break down the performance by language, which revealed notable heterogeneity.

High-Resource Languages: Languages such as Bengali, Marathi, and Urdu, which benefit from stronger NLP infrastructure and larger speaker populations, consistently received the highest human ratings (e.g., Bengali mean score 4.7 with 100% ≥ 3). LLM-as-a-judge based scores, while lower, followed similar trends, reflecting relatively better model performance in well-supported languages.

Mid-Resource Languages: Languages including Nepali, Sinhala, and Telugu demonstrated moderate performance. Interestingly, Telugu received higher scores from Prometheus than human annotators, possibly due to differences in model evaluation heuristics or familiarity with linguistic features.

Low-Resource Languages: Languages like Dzongkha, Pashto, and Dhivehi were rated low by both humans and the automated judge, reflecting persistent challenges in modeling fluency, cultural nuances, and data scarcity inherent to these languages.

5.5 Semantic Categorization and Cultural Tagging Consistency

Instruction Categorization: We evaluated DeepSeek R1's ability to classify Alpaca instructions into three categories: General, Language-related, and Culture-specific by comparing its predictions against our ground truth human labels (All reported accuracy values reflect comparison against a human-annotated ground truth, created by a team of native reviewers fluent in multiple South Asian languages). Using Jaccard similarity and exact label matching, we found that the model achieved over 70% Jaccard agreement across all categories, with an overall exact label match of 88%. This high agreement demonstrates that the model closely mirrors human judgment for broad semantic classification, though some subtle cultural nuances remain challenging.

Cultural Element Tagging: Again, model labels were compared directly to human ground truth using exact match rates. In a more granular evaluation, the model exhibited high agreement with human labels on culturally rich instructions in clearly defined categories such as Food & Beverages (100%) and Entertainment (90%). However, categories with greater ambiguity or lower frequency, such as Science, had notably poor agreement (0%). The overall exact label consistency was 78%, indicating solid but imperfect performance in culturally nuanced semantic tagging (See Appendix A).

5.6 Comparative Localization Analysis Between SOTA-LLMs

A focused comparative study between Command R and Llama 4 Maverick on localization tasks for India and Sri Lanka highlighted key differences.

Command R Localized 92% of the 7,833 instruction-input-output triples, outperforming Llama 4 Maverick, which localized 85%. It also demonstrated greater efficiency with an average generation time of 0.8 seconds per instance compared to 1.5 seconds. Command R's outputs were broadly factually accurate and culturally comprehensive across locales.

Conversely, while **Llama 4 Maverick** localized fewer instances, it occasionally incorporated creative cultural expressions and nuanced responses. This creative flair, although less consistent, introduces a dimension of cultural expressiveness which we term "LLMs being Culturally Creative", a trait potentially valuable depending on application context.

6 Conclusion

This paper presents SAID, the first culturally grounded instruction corpus spanning 15 South-Asian low-resource languages, 8 SAARC countries, and 10 cultural domains. Built via automated filtering, native-expert validation, and country-specific localization, SAID lets LLMs learn authentic linguistic and cultural nuance. LoRA fine-tuning on SAID produces SAID LLaMA 3.1 Instruct, which human and LLM judges rate as substantially more culturally aligned, factually accurate, and fluent than the base model.

Beyond culture-specific prompts, the same model generalises to standard NLP tasks such as translation, summarisation, and QA as it shows consistent gains over the base model—demonstrating that cultural adaptation does not come at the expense of general capability.

We will release all datasets and resources under an open-access Creative Commons CC BY 4.0 license, supporting the broader research community and advancing the equitable development of NLP technologies for South Asian languages.

Limitations

Despite the comprehensive scope of our study, several limitations remain. First, we were unable to perform detailed spot-check error analyses on localized outputs. Specifically, categorizing and quantifying error types such as mis-localization, unidiomatic phrasing, and factual inaccuracies on a representative sample (e.g., 100 instances per country) was beyond our current resources. This limits granular insights into specific challenges faced during localization.

Second, our fine-tuning experiments were conducted solely on the LLaMA 3.1 Instruct model. We did not evaluate the effectiveness of the SAID on a few other LLMs, which constrains the generalizability of our findings across different architectures or training paradigms. Nevertheless, our primary objective is to establish the impact of culturally-adaptive datasets on SOTA architectures as a proof of concept. For this reason, we fine-tuned one of the most recent LLMs using the proposed dataset.

Third, due to computational and resource constraints, we were unable to translate the entire localized SAID-English instruction dataset into all 15 target languages or conduct fine-tuning on these fully translated corpora. As a result, the full potential of culturally adaptive fine-tuning across the entire multilingual dataset remains unexplored.

Future work will prioritize addressing these gaps by performing systematic error type analyses, extending fine-tuning experiments to diverse LLM architectures, and fully translating and leveraging the complete SAID dataset to unlock deeper cultural understanding and performance improvements across South Asian languages.

Acknowledgments

We sincerely thank all the native speakers and domain experts from the South Asian region who contributed their invaluable time and expertise to the data annotation, cultural validation, and evaluation processes.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

HWK Aravinda, Rashad Sirajudeen, Samith Karunathilake, Nisansa de Silva, Surangika Ranathunga, and Rishemjit Kaur. 2025. Sinllama–a large language model for sinhala. *arXiv preprint arXiv:2508.09115*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert. *arXiv preprint arXiv:2304.11434*.

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.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M Khapra, Raj Dabre, Rudra Murthy, Anoop Kunchukuttan, and 1 others. 2024. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint arXiv:2401.15006*.

Deepanway Ghosal, Yew Ken Chia, Navonil Majumder, and Soujanya Poria. 2023. Flacuna: Unleashing the problem solving power of vicuna using flan fine-tuning. *arXiv preprint arXiv:2307.02053*.

Yuting Guo, Anthony Ovadje, Mohammed Ali Al-Garadi, and Abeed Sarker. 2024. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, 31(10):2181–2189.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Dorit Hadar-Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. 2024. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using schwartz's theory of basic values. *JMIR Mental Health*, 11:e55988.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2024. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *arXiv preprint arXiv:2409.12958*.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, and 1 others. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Ayana Niwa and Hayate Iso. 2024. Ambignlg: Addressing task ambiguity in instruction for nlg. *arXiv preprint arXiv:2402.17717*.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Soham Parikh, Quaizar Vohra, Prashil Tumbade, and Mitul Tiwari. 2023. Exploring zero and few-shot techniques for intent classification. *arXiv preprint arXiv:2305.07157*.

Parth Patwa, Simone Filice, Zhiyu Chen, Giuseppe Castellucci, Oleg Rokhlenko, and Shervin Malmasi. 2024. Enhancing low-resource llms classification with peft and synthetic data. *arXiv preprint arXiv:2404.02422*.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *arXiv preprint arXiv:2502.16534*.

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024a. Translating across cultures: Llms for intralingual cultural adaptation. *arXiv preprint arXiv:2406.14504*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, and 1 others. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Bibek Upadhayay and Vahid Behzadan. 2023. Taco: Enhancing cross-lingual transfer for low-resource languages in llms through translation-assisted chain-of-thought processes. *arXiv preprint arXiv:2311.10797*.

Sowmya Vajjala and Shwetali Shimangaud. 2025. Text classification in the llm era–where do we stand? *arXiv preprint arXiv:2502.11830*.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. 2024. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? *Preprint*, arXiv:2402.13703.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *arXiv preprint arXiv:2304.14402*.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2024. Self-pluralising culture alignment for large language models. *arXiv preprint arXiv:2410.12971*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024a. Lingualift: An effective two-stage instruction tuning framework for low-resource language tasks. *arXiv preprint arXiv:2412.12499*.

Shaolei Zhang, Kehao Zhang, Qingkai Fang, Shoutao Guo, Yan Zhou, Xiaodong Liu, and Yang Feng. 2024b. Bayling 2: A multilingual large language model with efficient language alignment. *arXiv preprint arXiv:2411.16300*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

A Appendix

A.1 South Asia and SAARC Member Countries

South Asia is a culturally and linguistically diverse region, encompassing eight countries: Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. These nations are united under the South Asian Association for Regional Cooperation (SAARC), established in 1985 to promote regional cooperation and development.

Selected Languages and Their Distribution The following 15 languages were selected for this study based on their prevalence, cultural significance, and representation across the SAARC member countries:

Language	Countries Spoken-In
Sinhala	Sri Lanka
Nepali	Nepal, India (Sikkim, Darjeeling)
Maithili	India (Bihar, Jharkhand), Nepal
Punjabi	India (Punjab), Pakistan (Punjab)
Assamese	India (Assam)
Sanskrit	India
Urdu	Pakistan, India
Bengali	Bangladesh, India (West Bengal)
Dhivehi	Maldives
Pashto	Afghanistan, Pakistan
Dari	Afghanistan
Awadhi	India (Uttar Pradesh, Madhya Pradesh)
Marathi	India (Maharashtra)
Telugu	India (Andhra Pradesh, Telangana)
Dzongkha	Bhutan

Table 2: Selected South Asian languages and their geographical distribution across SAARC countries.

Language Categorization The selected languages were categorized into high-resource, mid-resource, and LRLs based on factors such as the availability of digital resources, linguistic research, and computational tools:

High-Resource Languages: Languages with extensive digital resources, research, and computational tools available. This group includes Bengali, Hindi, Punjabi, Urdu, Marathi, and Telugu.

Mid-Resource Languages: Languages with moderate digital resources and computational tools. This group includes Nepali, Pashto, Dari, Assamese, Awadhi, and Dzongkha.

Low-Resource Languages: Languages with limited or no digital resources and computational tools. This group includes Sinhala, Maithili, Sanskrit, and Dhivehi.

A.2 Initial Instruction Categorization

We began with the Alpaca dataset (Taori et al., 2023), comprising 52,000 instruction instances. Using DeepSeek R1 as a zero-shot classifier, we categorized instructions into three classes: **Language-related** (grammar, translation, text manipulation), **Culture-relevant** (books, places, people, traditions, etc.), and **General** (instructions not fitting the first two categories). The classification prompt was:

> Follow these steps to classify the instruction:
>
> Is it about language (grammar, translation, text manipulation)? → Language-related Instruction
>
> If not, is it about culture (books, places, people, traditions)? → Culture-relevant
>
> If neither, → General.
>
> Instruction: [Insert Instruction Here]
>
> Final Category: [Your answer]

The resulting distribution was: General (28,800), Culture-relevant (5,300), and Language-related (17,400) instructions.

A.3 Semantic Cultural Tagging

Next, we refined the culture-relevant and language-related subsets by categorizing them into thematic cultural elements. After removing the *science* category, the final categories were:

- **Literature**: story, poem, novel, essay, journalism, poetry, dictionary, etc.
- **Entertainment**: movie, song, game, performance, music, art, fashion, concert
- **Geography**: country, city, place, region, map, travel, tourism
- **People**: career, profession, biography, historical figures, personality, family
- **History**: historical, battle, war, empire, ancient, era, period, old
- **Flora & Fauna**: animal, species, wildlife, trees, plants, endemic, creature
- **Sports**: sport, athlete, game, competition

- **Festivals**: festival, celebration, holiday, tradition, religion, event, function, show
- **Food & Beverages**: restaurant, cuisine, food, meal, beverages, alcohol, taste, drink, cook
- **Other**: Uncategorized or ambiguous instances

We ran DeepSeek R1 with the following prompt template to assign each instruction to one category:

> Given the following instruction and its response, classify it into one of these categories: literature, entertainment, geography, people, history, flora&fauna, sports, festivals, food&bev, other.
>
> Consider these keywords for each category: `[Insert category keywords here]`.
>
> Instruction: [Instruction text]
>
> Input: [Input text]
>
> Output: [Output text]
>
> Respond with ONLY the category name, nothing else.

A.4 Category Distribution and Manual Correction

The categorized instances were distributed as follows:

Category	Number of Instances
Literature	8,356
Other	4,173
Entertainment	1,902
People	1,679
Geography	1,372
History	936
Food & Beverages	810
Flora & Fauna	582
Sports	264
Festivals	138

Subsequently, the literature and other categories were manually reviewed to identify and reassign instances with clearer relations to other categories. Human annotators corrected mislabeling, especially for borderline cases. For example, machine learning book-related instructions were classified under *science* by the LLM but retained that label since they were considered general topics difficult to localize culturally and thus excluded from South Asian cultural adaptation.

> *Example:*
>
> { "instruction": "Generate a list of five books related to machine learning.", ... "cultural_element": "science" }

This human-in-the-loop approach ensured cultural relevance and semantic correctness throughout the dataset.

A.5 South Asian Instruction Dataset (SAID)

We named our dataset the **South Asian Instruction Dataset (SAID)**. The finalized dataset, derived from the Alpaca dataset and suitable for South Asian country-wise cultural adaptation, contains 7,833 instances, which we call the `culture_alpaca_dataset`. The total instances and their distribution by cultural element are:

- Language & Grammar: 2,617 (33.4%)
- Literature: 1,224 (15.6%)
- Geography: 1,041 (13.3%)
- Entertainment: 985 (12.6%)
- History and Religion: 772 (9.9%)
- People: 612 (7.8%)
- Food & Beverages: 289 (3.7%)
- Sports: 115 (1.5%)
- Festivals: 92 (1.2%)
- Flora & Fauna: 86 (1.1%)

Examples of instances from the dataset include:

```
{ "instruction":"Generate a list of
adjectives related to a summer garden.",
"input":"",
"output":"Lush, verdant, vibrant,
wild, abundant, blooming, colorful,
fragrant, buzzing, sweltering, healthy,
humid, teeming, sunny, warm, peaceful,
tranquil, serene, fresh, flourishing.",
"label":"Language-related",
"cultural_element":"flora&fauna"
}
```

```
{ "instruction":"Name five countries
with the most number of UNESCO World
Heritage sites.",
"input":"",
"output":"1. Italy (54 sites) 2. China
(52 sites) 3. Spain (48 sites) 4. France
(45 sites) 5. Germany (43 sites)",
"label":"Culture-relevant",
"cultural_element":"geography"
}
```

A.6 Cultural Adaptation for SAARC Countries

We performed cultural adaptation of the `culture_alpaca_dataset` to the eight SAARC countries using LLMs with few-shot prompt engineering, beginning with Sri Lanka. We selected Command R and Llama 4 Maverick due to their recent development, multilingual capabilities, and support for low-resource languages, making them suitable for this task.

The localization prompt used for Sri Lanka was:

```
You're an AI assistant converting
general English instruction data
into culturally relevant instructions
for Sri Lanka.
IMPORTANT: Always provide responses
in English only, not in Sinhala,
Tamil, or other Sri Lankan languages.
Make the content culturally relevant
and exclusive to Sri Lanka while
keeping it accessible to English speakers.
Also ensure that any names, places,
or cultural references in both
instruction and input are
localized to Sri Lanka.

Given:
Instruction: {instruction}
Input: {input_text}
Output: {output_text}

Rewrite the instruction, input (if present),
and output to make them contextually
and culturally appropriate for Sri Lanka.
The response should be similar in
style to these examples:
1. Instruction: "Provide one example
for a cultural practice in Sri Lanka."
   Input: "Traditional festivals"
   Output: "An example of a cultural
   practice is the Sinhala and Tamil
   New Year celebration, a major cultural
   festival marking the end of the harvest
   season."
2. Instruction: "Describe a historical
figure from Sri Lanka."
   Input: "A notable leader"
   Output: "King Parakramabahu I was one
   of Sri Lanka's greatest monarchs
   who unified the country and built
  massive irrigation systems in the 12th century...

Your response MUST be
a valid JSON object with
exactly this format:
{
  "instruction": "your English
  instruction here",
  "input": "your Sri Lanka-localized
  input here (if original input exists)",
  "output": "your English output here"
}
```

A sample localized instance for Sri Lanka is:

```
{ "instruction": "Compare and contrast
the 3 most popular sports in Sri Lanka.",
"input": "",
"output": "Cricket, volleyball, and
rugby are among the popular sports in
Sri Lanka. Cricket is highly popular
and emphasizes teamwork, strategy,
and skill, with a significant focus
on batting, bowling, and fielding.
Volleyball, on the other hand, is a team
sport that relies heavily on agility,
quick reflexes, and coordination among
team members. Rugby, similar to cricket
in its team aspect, demands strength,
physical endurance, and tactical play.
Overall, all three sports require
dedication, teamwork, and a certain
level of physical fitness, highlighting
the importance of these qualities in Sri
Lankan sports culture.",
"label": "Culture-relevant",
"cultural_element": "sports",
"country": "Sri Lanka"
}
```

We engaged a team of eight native reviewers, each fluent in multiple South Asian languages, to fact-check and rate the quality of the country-wise localized instances on a scale from 1 to 10. The average quality ratings per country were:

- Sri Lanka: 10/10
- India: 10/10
- Nepal: 9/10
- Pakistan: 10/10
- Afghanistan: 8/10
- Bhutan: 7/10
- Bangladesh: 9/10
- Maldives: 7/10

A.7 Heatmap of Mean Ratings by Language

Language	Human Mean	Human%3	LLM Mean	LLM%3
Bengali	4.70	100%	2.20	20%
Marathi	3.40	90%	2.90	70%
Urdu	3.50	70%	1.70	25%
Dari	3.10	80%	1.00	0%
Sinhala	2.70	40%	2.30	30%
Nepali	2.30	60%	1.80	10%
Telugu	2.10	30%	2.70	50%
Assamese	2.10	30%	2.20	40%
Panjabi	2.50	50%	1.20	0%
Awadhi	2.30	20%	1.90	10%
Dzongkha	1.50	20%	1.50	20%
Sanskrith	1.70	30%	1.30	10%
Pashto	1.56	11%	1.56	11%
Dhivehi	1.82	18%	1.91	27%
Maithili	2.40	60%	1.90	30%

Table 3: Human and LLM evaluation scores per language.

Insights: Languages with stronger NLP resources or larger user bases (e.g., Bengali, Marathi, Urdu) tend to have higher human ratings and sometimes higher LLM ratings, although LLM remains more conservative overall.

Smaller or lower-resource languages (Dzongkha, Pashto, Dhivehi, Sanskrith) tend to have lower scores from both humans and LLM, possibly reflecting challenges in cultural alignment or language fluency.

In several languages, the LLM judge is notably more critical than humans, emphasizing the difficulty of automated evaluation in culturally sensitive contexts.

In a few cases (e.g., Telugu), the LLM judge gives higher scores than humans, suggesting variability in evaluation criteria or model behavior.

A.8 Per-Element Agreement and Overall Label Consistency

We computed the per-element agreement using the Jaccard similarity coefficient to evaluate how well the LLM labeling aligned with human annotations across cultural categories. The results are as follows:

- Entertainment: 90.00%
- Festivals: 60.00%
- Flora & Fauna: 70.00%
- Food & Beverages: 100.00%
- Geography: 54.55%
- History/Religion: 61.54%
- Literature: 72.73%
- Other: 36.84%
- People: 83.33%
- Science: 0.00%
- Sports: 70.00%

The overall exact label consistency between human and LLM annotations was **78.00%**, indicating that the LLM exactly matched the human label in 78 out of 100 instances, which is considered relatively high for a multi-class cultural annotation task with nuanced categories.

Analysis of Per-Element Agreement: High agreement categories:

- **Food & Beverages (100%):** Perfect overlap indicates very consistent categorization between humans and the LLM.
- **Entertainment (90%) and People (83.33%):** Strong agreement suggests these categories are well-defined and clearly distinguishable by both human annotators and the model.
- **Sports (70%) and Flora & Fauna (70%):** Reasonably good agreement, indicating moderate clarity in category boundaries.

Moderate agreement categories:

- **Literature (72.73%) and History/Religion (61.54%):** Reasonable agreement, with some disagreements likely due to subtle distinctions or ambiguous instructions.
- **Festivals (60%) and Geography (54.55%):** Moderate agreement showing some confusion or overlap with other categories, possibly due to shared cultural references.
- **Other (36.84%):** Low agreement is expected, as this category often captures ambiguous or outlier instances leading to subjective interpretations.

Low agreement category:

- **Science (0%):** No overlap observed, suggesting either very few instances labeled as science or consistent misclassification by the model. This indicates a weakness in the model's understanding or the ambiguity of scientific instructions within this dataset.

The 22% of mismatches primarily originate from categories with moderate to low agreement, such as "other," "science," and "festivals," reflecting the inherent ambiguity or overlapping semantics within these cultural categories.

A.9 Cross-Regional Cultural Transferability Testing

In HRLs, the fine-tuned model consistently outperformed the base model, delivering responses that were more concise, contextually relevant, and culturally accurate, effectively bridging explicit and implicit localization gaps with scores between 4.5 and 5.

In contrast, LRLs exhibited substantially weaker performance, with outputs often lacking relevant South Asian cultural grounding, showing noise, or becoming off-topic altogether. This emphasizes the persistent challenges of cross-lingual and cultural transfer in languages with limited NLP resources or

divergent cultural frameworks. For each language, we posed 30 open-ended questions—five per language—about South Asian culture using both models to detect explicit and implicit localization gaps (Weber et al., 2024).

Some instances were machine-translated from the country-specific localized `alpaca_culture_dataset`, excluding those in the SAID-English-minor dataset.

Lang	BM Avg.	FM Avg.	Exp	Imp
French	3.0–4.5	4.5–5.0	No	Minor
Arabic	3.5–4.5	4.0–4.5	Minor	Minor
Chinese	4.0–4.5	4.5–5.0	Minor	Minor
Uyghur	N/A	N/A	Large	Large
Hawaiian	N/A	N/A	Large	Large
Zulu	N/A	N/A	Large	Large

Table 4: Cross-regional evaluation results comparing base and finetuned models on South Asian cultural questions across non-South Asian languages. From left to right: Lang - Language, BM Avg.- Base Model Average Score, FM Avg. - Finetuned Model Average Score, Exp - Explicit Gap, Imp - Implicit Gap

A.10 Human Annotators and Domain Expert Profile

The human annotation and evaluation process was conducted by a team of undergraduate and graduate student volunteers recruited from various South Asian countries. These annotators were carefully selected based on their proficiency in at least two of the 15 selected South Asian languages, ensuring linguistic competence and cultural familiarity essential for high-quality annotation.

Before beginning their tasks, all annotators received comprehensive guidance and training on the annotation protocols, cultural sensitivity, and quality standards to maintain consistency and reliability throughout the project.

Their voluntary contribution was invaluable in validating dataset quality and providing culturally grounded evaluation insights. Upon successful completion of their assignments, each annotator was formally acknowledged with a letter of appreciation, officially sealed by the research institution, recognizing their essential role and dedication to the project.

A.11 General Purpose Task Results

1) FLORES200 Multilingual Evaluation Benchmark Results

Language	Metric	Base	Tuned	Δ
Sinhala	BLEU	0.03	0.06	0.03
	NIST	0.48	1.03	0.55
	chrF	0.34	0.31	-0.03
	WER	7.69	1.36	6.33
Nepali	BLEU	0.02	0.12	0.10
	NIST	0.33	1.60	1.27
	chrF	0.33	0.48	0.16
	WER	11.73	1.20	10.53
Urdu	BLEU	0.03	0.11	0.09
	NIST	0.36	1.43	1.07
	chrF	0.32	0.44	0.12
	WER	9.97	3.13	6.84
Bengali	BLEU	0.08	0.13	0.05
	NIST	0.83	1.46	0.63
	chrF	0.46	0.44	-0.03
	WER	3.07	1.60	1.47
Pashto	BLEU	0.02	0.06	0.04
	NIST	0.29	0.74	0.45
	chrF	0.29	0.34	0.05
	WER	12.10	5.46	6.64
Dari	BLEU	0.02	0.16	0.13
	NIST	0.28	1.78	1.50
	chrF	0.32	0.49	0.18
	WER	12.41	2.18	10.23
Awadhi	BLEU	0.03	0.13	0.10
	NIST	0.29	1.65	1.36
	chrF	0.32	0.47	0.15
	WER	11.75	3.09	8.66
Marathi	BLEU	0.03	0.12	0.09
	NIST	0.35	1.62	1.27
	chrF	0.32	0.48	0.16
	WER	11.10	1.62	9.48
Telugu	BLEU	0.05	0.11	0.05
	NIST	0.69	1.35	0.66
	chrF	0.42	0.42	0.01
	WER	6.13	1.20	4.93
Dzongkha	BLEU	0.00	0.00	-0.00
	NIST	0.16	0.06	-0.09
	chrF	0.20	0.02	-0.18
	WER	12.02	1.06	10.96
Maithili	BLEU	0.02	0.16	0.14
	NIST	0.28	2.04	1.76
	chrF	0.29	0.51	0.22
	WER	12.02	1.26	10.76
Punjabi	BLEU	0.04	0.03	-0.01
	NIST	0.40	0.53	0.13
	chrF	0.35	0.20	-0.15
	WER	10.78	1.91	8.86
Assamese	BLEU	0.03	0.07	0.04
	NIST	0.50	1.04	0.54
	chrF	0.36	0.38	0.02
	WER	6.23	1.98	4.25
Sanskrit	BLEU	0.01	0.05	0.04
	NIST	0.24	1.19	0.96
	chrF	0.27	0.39	0.12
	WER	11.71	2.50	9.21

Table 5: FLORES200 multilingual evaluation: scores for the base versus SAID-fine-tuned models, plus absolute improvements (Δ).

2) Aya Evaluation Suite Results

Lang	Model	BLEU	METEOR	R-1	R-2	R-L	CoSim
Nepali	Base	0.12	0.18	0.00	0.00	0.00	0.11
Nepali	Finetuned	0.15	0.28	0.00	0.00	0.00	0.15
Bengali	Base	0.08	0.20	0.00	0.00	0.00	0.14
Bengali	Finetuned	0.08	0.18	0.00	0.00	0.00	0.13
English	Base	0.12	0.40	0.42	0.18	0.26	0.40
English	Finetuned	0.16	0.45	0.46	0.23	0.30	0.43
Hindi	Base	0.12	0.21	0.05	0.02	0.05	0.12
Hindi	Finetuned	0.21	0.34	0.07	0.02	0.07	0.17
Pashto	Base	0.05	0.15	0.06	0.05	0.06	0.22
Pashto	Finetuned	0.04	0.16	0.03	0.00	0.03	0.26
Sinhala	Base	0.02	0.12	0.04	0.02	0.03	0.10
Sinhala	Finetuned	0.02	0.12	0.05	0.02	0.05	0.09
Telugu	Base	0.03	0.13	0.03	0.00	0.03	0.09
Telugu	Finetuned	0.02	0.12	0.02	0.00	0.02	0.06
Urdu	Base	0.05	0.15	0.13	0.03	0.12	0.27
Urdu	Finetuned	0.04	0.16	0.16	0.03	0.14	0.32

Table 6: Aya summarization benchmark. Lang = target language, Model= system variant (Base vs. SAID-Finetuned); BLEU and METEOR are standard n-gram overlap metrics; R-1, R-2, and R-L are ROUGE-1/-2/-L F-scores; CoSim is TF-IDF cosine similarity. Higher values indicate better faithfulness and fluency.

3) L3Cube Question Answering Benchmark Results

Lang	Model	FA	Rel	Clar	LangC	Concise
Bengali	Base	1	1	1	2	2
Bengali	Finetuned	5	5	3	3	3
Hindi	Base	3	3	2	2	2
Hindi	Finetuned	5	5	5	5	4
Maithili	Base	5	5	5	4	4
Maithili	Finetuned	5	5	5	4	4
Nepali	Base	4	4	4	4	4
Nepali	Finetuned	5	5	5	5	5
Punjabi	Base	2	2	2	2	2
Punjabi	Finetuned	2	2	2	2	2
Telugu	Base	2	3	1	1	1
Telugu	Finetuned	3	3	3	3	3
Urdu	Base	4	4	4	3	3
Urdu	Finetuned	5	4	4	2	2

Table 7: L3Cube IndicNLP question-answering benchmark. **FA** = factual accuracy; **Rel** = relevance; **Clar** = clarity; **LangC** = language consistency; **Concise** = conciseness. Higher scores are better (1–5 scale).