Advancing NLP Equity: A Secondary Benchmark Evaluation of Multilingual Language Models for Underrepresented Languages

Md Muntaqim Meherab[1⋆], Salman[1], Md. Maruf Billah[2], Kazi Shakkhar Rahman[3], Liza Sharmin[4], Tanvirul Islam[1], Z N M Zarif Mahmud[1], Nuruzzaman Faruqui[1], Sheak Rashed Haider Noori[1], and Touhid Bhuiyan[4]

[1] Daffodil International University, Dhaka, Bangladesh
{meherab2305101354, salman2305101404, tanvirulislam.cse, zarif.cse, headcse, deanfhss, faruqui.swe}@diu.edu.bd
[2] North South University, Dhaka, Bangladesh
maruf.billah.232@northsouth.edu
[3] Islamic University of Technology, Gazipur, Bangladesh
shakkharrahman@iut-dhaka.edu
[4] Washington University of Science and Technology, Alexandria, VA, USA
touhid.bhuiyan@wust.edu

Abstract. Recent multilingual language models promise support for "100+ languages," yet speakers of Indigenous and other underrepresented languages still often do not see themselves in these advances. In this work, we take a deliberately simple, secondary-benchmark perspective: rather than proposing a new model or dataset, we re-evaluate an off-the-shelf multilingual natural language inference (NLI) model on public benchmarks that explicitly include Indigenous languages of the Americas. Concretely, we use the AmericasNLI benchmark for ten Indigenous languages and XNLI for English and Spanish, and we evaluate the widely used `joeddav/xlm-roberta-large-xnli` model under a fixed, zero-shot protocol. Our goal is to answer three questions: (*i*) How large is the performance gap between high-resource and underrepresented languages under the same model and task? (*ii*) Are these gaps consistent across languages, or do some communities fare systematically worse than others? (*iii*) What kinds of qualitative errors arise, and what do they suggest about cultural and linguistic mismatch? Our experiments reveal a striking discrepancy: while English and Spanish reach almost perfect accuracy on XNLI (around 99.8% on our runs), the same model averages only about 43% accuracy across ten Indigenous languages in AmericasNLI, with none exceeding 47%. We also show qualitative NLI failures in Quechua that point to difficulties with morphology, idioms, and discourse-level inference. We argue that even such a simple re-analysis can serve as a low-cost yet high-impact tool for making inequities in multilingual NLP visible, especially for communities that rarely appear in headline benchmarks.

⋆ Corresponding author: meherab2305101354@diu.edu.bd

Keywords: Multilingual NLP · Natural language inference · Low-resource languages · Indigenous languages · Language equity · AmericasNLI · XLM-R

1 Introduction

Large pretrained language models have reshaped natural language processing (NLP) in just a few years. Transformer architectures [13] and multilingual pretraining [4, 6] have led to impressive gains across tasks and languages, and many models now advertise support for "100+ languages" out of the box. From a distance, this can give the impression that the field is on track to serve a linguistically diverse world.

A closer look, however, tells a more uneven story. A growing body of work shows that most of the benefits of modern NLP are concentrated in a small set of high-resource languages, typically major European and East Asian languages [2, 9]. Speakers of Indigenous and minoritized languages often see little or no improvement in everyday tools—if such tools exist at all. This is not just a technical detail; it is part of a broader pattern in which linguistic communities already marginalized in education, media, and governance are left behind in the digital sphere as well.

In this paper, we focus on this gap from a pragmatic angle. Rather than introducing a new model, we ask a simpler question: *If we take a widely used multilingual model exactly as it is, and we evaluate it carefully on a benchmark that explicitly includes underrepresented languages, what story do the numbers tell?* This type of secondary benchmarking is not glamorous. However, it can be done quickly, it is reproducible, and it can provide clear evidence that is easy to communicate both inside and outside the research community.

We centre our study on **Natural Language Inference (NLI)**, a standard testbed for higher-level semantic understanding. We treat NLI as a proxy for whether a language model can handle non-trivial semantics in languages that are rarely represented in its training data. For languages with very limited NLP resources, even basic NLI competence is far from guaranteed.

Concretely, we evaluate the following setup:

- **Model:** `joeddav/xlm-roberta-large-xnli`, an XLM-R large model fine-tuned on XNLI for NLI in 15 languages.
- **Low-resource benchmark:** AmericasNLI, an NLI dataset in ten Indigenous languages of the Americas, designed for zero-shot evaluation of large multilingual models.
- **High-resource reference:** XNLI test sets for English and Spanish, two languages explicitly included during fine-tuning.

We deliberately *do not* fine-tune or adapt the model. Instead, we use a uniform, zero-shot inference protocol across all languages. This allows us to ask: under the same task, architecture, and label space, how differently does the model behave depending on which community's language it is exposed to?

We structure our analysis around three research questions:

RQ1: How well does a widely used multilingual NLI model perform on truly low-resource Indigenous languages compared with high-resource languages?
RQ2: Are performance gaps consistent across Indigenous languages, or do some communities experience systematically worse performance than others?
RQ3: What qualitative error patterns emerge, and what do they suggest about the model's handling of morphology, idioms, and discourse in these languages?

Our experimental design is intentionally modest, both in scope and in resource demands. We use only public datasets and a single off-the-shelf model. The entire evaluation can be reproduced on a single GPU within roughly a day, making it accessible to students, small labs, and community-based researchers who may not have access to large compute clusters.

Despite its simplicity, the results are sobering. In our runs, the XNLI-fine-tuned model reaches nearly perfect accuracy on English and Spanish (around 99.9% and 99.6%, respectively), but averages only about 43% accuracy across the ten AmericasNLI languages. Accuracy for individual Indigenous languages hovers in the low-40% range, with no language above 47%. In other words, a model that is effectively "solving" NLI in high-resource settings behaves more like a weak baseline when asked to support Indigenous communities.

Beyond the headline numbers, qualitative error analysis for Quechua reveals systematic confusions between contradiction and neutral, as well as a tendency to over-predict entailment in the presence of complex morphology and discourse markers. These errors align with broader concerns that current multilingual models often underrepresent the structural and cultural diversity of the world's languages [1, 9].

Our contributions are:

- A simple but rigorous secondary-benchmark evaluation of a widely used multilingual NLI model on AmericasNLI and XNLI, quantifying the performance gap between high-resource and Indigenous languages.
- A set of language-level and aggregate metrics that make these gaps easy to communicate and compare.
- A small but concrete qualitative analysis of NLI failures in Quechua, illustrating how errors connect to morphology, idioms, and discourse.
- A reproducible, low-cost evaluation pipeline that can serve as a template for similar audits of other multilingual models and tasks.

Our results do not claim to be the final word on fairness in multilingual NLI. They are, however, a reminder that even small, focused re-evaluations can surface inequities that might otherwise be hidden behind impressive average scores.

2 Background and Related Work

2.1 Multilingual Pretrained Language Models

Transformer-based pretrained language models such as BERT [6] and its multilingual variants have become the backbone of modern NLP. Multilingual BERT

(mBERT) and XLM-R [4] extend this paradigm by jointly pretraining on large corpora spanning dozens or hundreds of languages. These models have delivered strong cross-lingual transfer on tasks such as part-of-speech tagging, question answering, and NLI, especially for languages with reasonable amounts of text in pretraining corpora.

However, pretraining coverage is not synonymous with equitable performance. Even when a language is technically present in the pretraining mix, it may be represented by orders of magnitude fewer tokens than English or other high-resource languages. Recent studies have shown that, in practice, performance tends to track data availability and socio-economic status of language communities rather than any intrinsic linguistic property [2]. To measure this imbalance more systematically, benchmarks such as XTREME [8] and its successor XTREME-R [12] were developed to span dozens of languages and several task types, and in doing so they exposed how sharply cross-lingual transfer quality degrades once a target language falls outside the well-resourced core of the pretraining corpus. Lauscher et al. [10] pushed this finding further, showing that zero-shot transfer is considerably more brittle than headline numbers suggest—even modest shifts in morphological complexity or domain can cause substantial accuracy drops for languages at the lower end of the resource spectrum.

2.2 Cross-Lingual NLI Benchmarks

Natural Language Inference has become a standard testbed for semantic understanding across languages. XNLI [5] extends the English MultiNLI dataset to 15 languages via translation, providing a benchmark for evaluating multilingual sentence representations and cross-lingual transfer. Models like XLM-R are often fine-tuned on XNLI for these languages and then reused for zero-shot classification in other settings.

While XNLI has driven substantial progress, its language coverage remains skewed toward high-resource languages. To address this gap, AmericasNLI introduces NLI test and validation sets for ten Indigenous languages of the Americas, including Asháninka, Aymara, Bribri, Guaraní, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, and Wixarika. Ebrahimi et al. [7], who introduced AmericasNLI, designed it specifically to expose the limits of zero-shot transfer for languages that had been all but invisible in mainstream multilingual evaluation—none of the target languages appear in the fine-tuning data of the models the benchmark is intended to test. Complementary evidence from Clark et al. [3] in the question-answering domain reinforces the same concern: their TyDi QA benchmark, built around typologically diverse languages, shows that model performance closely tracks the linguistic distance between a target language and the high-resource languages that dominate pretraining, making typological coverage a concrete fairness issue rather than a purely academic one. Our work builds directly on this line of research. Rather than training new models, we use AmericasNLI as a lens to examine how an off-the-shelf XNLI-fine-tuned model behaves when evaluated on these languages alongside its original high-resource targets.

2.3 Inequalities in Language Technology

Several studies have documented systematic inequalities in language technology. Joshi et al. [9] quantify how a small set of languages dominate NLP research and resources, while most of the world's languages receive little or no attention. Blasi et al. [2] provide a global perspective on disparities in performance across languages, linking them to a combination of resource availability, economic indicators, and colonial history.

Other work has argued that resource labels such as "low-resource" can obscure important distinctions between standardised, local, and contact languages, and can carry their own power dynamics [1]. Nekoto et al. [11] offer a practical illustration of this point through a large-scale, community-led machine translation initiative for African languages, demonstrating that sustainable progress for underserved communities depends not just on model improvements but on genuine participatory involvement of those communities in data creation, evaluation, and deployment decisions. From a governance perspective, these disparities have implications for who gets to shape the development and deployment of language technologies, and whose values and communicative practices are encoded in them.

Our study sits within this broader conversation but with a narrower empirical focus. We ask a concrete question about one widely used multilingual NLI model and two well-defined benchmarks, with the aim of providing clear, reproducible evidence that can be situated within a larger body of work on linguistic equity in NLP.

2.4 Secondary Benchmarking and Model Audits

There is a growing recognition that evaluating existing models under new conditions can be as valuable as building new models, especially for fairness and governance. Secondary benchmarking—reusing public models and datasets to answer new questions—offers an accessible way to conduct such audits. It lowers the entry barrier for researchers and community members who may not have resources to train large models but do have important questions about how those models behave.

In the multilingual setting, secondary benchmarking has been used to examine gender bias, toxicity, and domain shifts across languages. Our work follows this spirit: we use standard tools (Hugging Face Transformers), public benchmarks (AmericasNLI, XNLI), and a simple evaluation pipeline to highlight how much performance can change when moving from well-served to underrepresented languages.

Table 1 summarises how our study sits in relation to key strands of prior work on multilingual models, low-resource NLI, and linguistic equity.

Table 1: Positioning Our Work Within Prior Research

Area	**What Existing Work Shows**	**What Our Study Adds**
Multilingual pretrained language models	Large models like mBERT and XLM-R deliver strong results for well-represented languages, but performance tends to follow data availability and global socio-economic patterns.	We provide a concrete, task-specific demonstration of this imbalance by quantifying how sharply performance drops for Indigenous American languages under identical evaluation conditions.
Cross-lingual NLI benchmarks	XNLI has become a standard benchmark but is skewed toward high-resource languages; AmericasNLI was created to fill this gap by providing truly low-resource test sets.	Instead of proposing new data or models, we treat AmericasNLI as a diagnostic tool, using it to reveal how a commonly deployed XNLI-tuned model behaves when confronted with languages outside its design scope.
Inequalities in language technology	Prior studies document global disparities in NLP performance, resource availability, and historical factors shaping linguistic representation.	We offer a focused, reproducible case study showing how these systemic patterns manifest in a familiar task (NLI), turning abstract inequality findings into specific, measurable outcomes.
Secondary benchmarking and model audits	Auditing models under new conditions has become increasingly recognized as a lightweight but meaningful fairness practice.	We contribute a straightforward audit procedure that others can reuse, emphasizing accessibility for researchers without large computational budgets.
Evaluation of under-represented languages	Much prior work focuses on building models, datasets, or adaptation strategies for low-resource languages.	Our role is complementary: we highlight the baseline reality of out-of-the-box model behavior, establishing a reference point that any future improvements should exceed.

3 Task and Problem Setting

3.1 Natural Language Inference as a Probe

We frame our study around **Natural Language Inference (NLI)**. Given a *premise* sentence and a *hypothesis* sentence, the task is to predict whether the hypothesis is *entailed* by the premise, *contradicted* by it, or *neutral*. Even though NLI is an idealised task, it requires models to engage with semantic phenomena such as negation, quantification, lexical relations, and basic world knowledge.

For high-resource languages like English, NLI benchmarks have become mature enough that large pretrained models approach human-level performance. This makes NLI a useful probe: if a model that "solves" NLI in English performs poorly on the same task in an Indigenous language, this suggests that the underlying semantic capabilities have not transferred equitably.

Formally, let $\mathcal{L}$ denote the set of sentences in a given language, and let $\mathcal{Y} = \{\text{entailment}, \text{neutral}, \text{contradiction}\}$ be the label space. Each NLI example is a pair $(p, h) \in \mathcal{L} \times \mathcal{L}$ consisting of a premise p and a hypothesis h, together with a gold label $y \in \mathcal{Y}$. A classifier f_θ (here, our fine-tuned multilingual model) implements a mapping

$$f_\theta : \mathcal{L} \times \mathcal{L} \to \mathcal{Y}, \quad f_\theta(p, h) = \hat{y}, \tag{1}$$

where $\hat{y}$ is the predicted NLI label. In this work, we do not update θ; we evaluate a fixed pretrained f_θ in a zero-shot setting across languages.

3.2 Underserved Communities and Indigenous Languages

We follow prior work in treating speakers of Indigenous languages of the Americas as a concrete example of underserved communities in NLP. These languages are often underrepresented in educational systems, under-resourced in terms of digital tools, and sometimes endangered in terms of intergenerational transmission. For many of them, even basic NLP infrastructure—tokenizers, morphological analyzers, or spell-checkers—is still in early stages or entirely absent.

Importantly, these languages are not just data points or test cases. They encode rich cultural histories, oral traditions, and epistemologies that have developed over centuries. When language technologies fail to support these languages, they risk reinforcing existing patterns of exclusion from digital spaces. Our goal in this paper is not to "fix" these issues, but to help document them in a way that is concrete, reproducible, and legible to the broader NLP community.

3.3 Research Questions

Given this context, our study is guided by three research questions (RQ1–RQ3), which we restate here with slightly more operational detail:

RQ1: *How large is the performance gap between high-resource and Indigenous languages on NLI under a fixed multilingual model and evaluation protocol?* We measure language-wise accuracy and macro-F1, and compare the mean performance of English and Spanish (high-resource languages included in XNLI fine-tuning) against that of the ten AmericasNLI languages.

RQ2: *Do performance gaps vary substantially across Indigenous languages?* We examine per-language metrics to see whether all languages are equally disadvantaged, or whether some languages consistently receive better or worse support from the model.

RQ3: *What kinds of qualitative errors does the model make in an Indigenous language, and what do these errors suggest about its handling of morphology, idioms, and discourse?*
We focus on Quechua as a case study, collecting a small sample of misclassified NLI pairs and analysing where the model tends to over-predict entailment or confuse contradiction with neutral.

These questions are intentionally narrow. They do not address all dimensions of fairness or usability, but they provide a concrete starting point that others can extend with more sophisticated analyses or additional models.

4 Methodology

Our methodology is designed around a few principles: use only public resources; keep the evaluation protocol simple and transparent; and separate high-resource and low-resource settings as clearly as possible.

4.1 Datasets

AmericasNLI is an NLI benchmark that extends the XNLI framework to ten Indigenous languages of the Americas. For each language, it provides development and test splits with a few hundred examples each, derived from parallel annotations aligned with English NLI data. The target languages include:

Asháninka (`cni`), Aymara (`aym`), Bribri (`bzd`), Guaraní (`gn`), Nahuatl (`nah`), Otomí (`oto`), Quechua (`quy`), Rarámuri (`tar`), Shipibo-Konibo (`shp`), and Wixarika (`hch`).

In our experiments, we use the test split for each language. The number of test examples per language ranges from roughly 738 to 750, following the statistics provided with the dataset. We treat these languages as representative examples of underrepresented communities in current NLP practice: they are typologically diverse, historically marginalized, and largely absent from mainstream training corpora.

XNLI For a high-resource reference point, we use the XNLI test sets for English and Spanish. XNLI extends MultiNLI to 15 languages via translation and serves as a standard benchmark for cross-lingual NLI. Crucially, the

`joeddav/xlm-roberta-large-xnli` model we evaluate has been fine-tuned on XNLI for a subset of these languages, including English and Spanish. We treat performance on these two languages as an approximate upper bound for what the model can achieve under our evaluation protocol.

4.2 Model

We evaluate a single off-the-shelf model:

- **Model:** `joeddav/xlm-roberta-large-xnli`
- **Base architecture:** XLM-R large [4]
- **Fine-tuning:** Trained on XNLI for NLI in 15 languages

XLM-R itself is a multilingual masked language model pretrained on over 100 languages with the RoBERTa-style objective.
The `joeddav/xlm-roberta-large-xnli` checkpoint adds a classification head and fine-tuning on XNLI, making it a widely used backbone for zero-shot cross-lingual NLI and text classification.

This model is a natural choice for our study for two reasons. First, it is popular in practice; many applied systems rely on it, often under the assumption that it "supports" a wide range of languages out of the box. Second, its fine-tuning data explicitly excludes the Indigenous languages in AmericasNLI, making it a realistic example of a model that is powerful in high-resource contexts but not intentionally designed for the underrepresented languages we care about.

4.3 Evaluation Protocol

We adopt a uniform, zero-shot evaluation protocol across all languages and datasets. There is no fine-tuning or adaptation on AmericasNLI.

Input encoding. For each (premise, hypothesis) pair, we use the Hugging Face XLM-R tokenizer to encode the sentence pair with a maximum sequence length of 128 tokens. We rely on the standard sentence-pair encoding format used in `transformers`, which internally handles segment embeddings as appropriate for XLM-R.

Model inference. We run batched inference using a batch size of 32 and single-GPU acceleration. For each batch, we obtain the model's logits over the three NLI labels (entailment, neutral, contradiction). We then take the argmax to obtain the predicted class for each example.

Metrics. For each language and dataset, we report two standard classification metrics:

- **Accuracy**: the proportion of test examples for which the predicted label matches the gold label.

– **Macro-F1**: the unweighted average F1 score across the three classes, which helps reduce the impact of any label imbalance.

To support RQ1 and RQ2, we also compute aggregate mean accuracy for two groups:

– **High-resource group**: the English and Spanish XNLI test sets.
– **Low-resource group**: the ten AmericasNLI languages.

Formally, given a test set $\{(p_i, h_i, y_i)\}_{i=1}^{N}$ for a particular language, with gold labels $y_i \in \mathcal{Y}$ and model predictions $\hat{y}_i = f_\theta(p_i, h_i)$, the accuracy is

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}, \tag{2}$$

where $\mathbf{1}\{\cdot\}$ is the usual indicator function.

Let $\mathcal{C} = \{\text{entailment}, \text{neutral}, \text{contradiction}\}$ denote the set of classes. For each class $c \in \mathcal{C}$, we define precision P_c and recall R_c in the standard way from true positives, false positives, and false negatives. The macro-averaged F1 score is then

$$\text{Macro-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{2 P_c R_c}{P_c + R_c + \varepsilon}, \tag{3}$$

where ε is a small constant to avoid division by zero.

Error analysis. For RQ3, we focus on Quechua (`quy`) as a case study. We run the same inference procedure and then collect a small sample of misclassified examples. For each, we record the premise, hypothesis, gold label, and predicted label, and we manually inspect them to identify recurring patterns such as:

– Confusions between contradiction and neutral.
– Systematic over-prediction of entailment.
– Apparent difficulties with morphology or idiomatic expressions.

4.4 Implementation Details

All experiments are implemented in Python using the Hugging Face `datasets` and `transformers` libraries. We load AmericasNLI and XNLI directly from Hugging Face, and we rely on the model hub to fetch the `joeddav/xlm-roberta-large-xnli` checkpoint. We fix a random seed for data loading to ensure reproducibility, although the evaluation itself is deterministic given the trained model.

The entire pipeline—from data loading to metric computation—can be run in a single Colab notebook with GPU support.

5 Experiments and Results

In this section, we report our zero-shot evaluation of XLM-R_{large} (fine-tuned on XNLI) on the AmericasNLI benchmark and the XNLI test sets for English and Spanish. We focus on three aspects: (i) the overall performance gap between high- and low-resource languages, (ii) variation across Indigenous languages, and (iii) qualitative error patterns that shed light on cultural and morpho-syntactic mismatches.

5.1 Experimental Setup

We follow the protocol outlined in Section 4. For each language in AmericasNLI, we use the official test split and run the pretrained `joeddav/xlm-roberta-large-xnli` model in a pure zero-shot setting, without any further fine-tuning or adaptation. For XNLI, we evaluate on the English and Spanish test sets to approximate an upper bound for high-resource, in-distribution languages.

Unless otherwise noted, we use a maximum sequence length of 128 subword tokens and a batch size of 32. Longer sequences are truncated from the end. We compute accuracy and macro-averaged F1 (*macro-F1*) over the three NLI labels (entailment, neutral, contradiction). All predictions are obtained with a single forward pass per example.

5.2 Per-language Performance on AmericasNLI

Table 2 shows per-language performance on AmericasNLI. Accuracy ranges from roughly 40% to 47% across the ten Indigenous languages, with macro-F1 tracking accuracy closely. In other words, even the best-performing languages are just a bit above the random baseline of 33.3% for a balanced three-way classification task.

The model does slightly better on Shipibo-Konibo (0.467 accuracy, 0.456 macro-F1) and Nahuatl (0.463 / 0.446), and slightly worse on Rarámuri (0.403 / 0.352) and Wixarika (0.407 / 0.362). Overall, the spread between the best and worst Indigenous languages is modest (about six percentage points of accuracy), and the entire band is far from what would normally be considered "usable" performance in high-stakes applications.

Figure 1 visualizes this distribution. The bar plot makes the story easy to see at a glance: all Indigenous languages cluster just above chance level, with none approaching high-resource performance.

5.3 High- vs. Low-resource Comparison

To understand how large the gap is between high- and low-resource languages for the *same* model and *same* task, we compare AmericasNLI results with XNLI results for English and Spanish (Table 3). On XNLI, the model achieves almost

Language	Code	# Test	Acc.	Macro-F1
Aymara	aym	750	0.421	0.378
Bribri	bzd	750	0.443	0.419
Asháninka	cni	750	0.416	0.390
Guaraní	gn	750	0.460	0.428
Wixarika	hch	750	0.407	0.362
Nahuatl	nah	738	0.463	0.446
Otomí	oto	748	0.434	0.407
Quechua	quy	750	0.417	0.388
Shipibo-Konibo	shp	750	0.467	0.456
Rarámuri	tar	750	0.403	0.352

Table 2: Zero-shot NLI performance of XLM-R_{large} on AmericasNLI. Accuracy and macro-F1 are computed over three labels: entailment, neutral, and contradiction.

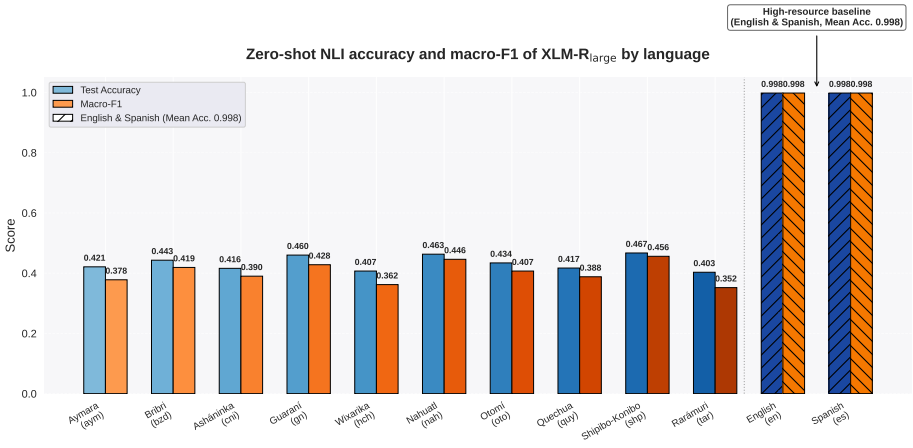

Fig. 1: Zero-shot accuracy of XLM-R_{large} on each AmericasNLI language. All languages sit in a narrow band around 43% accuracy, only slightly above the random baseline of 33.3%.

perfect accuracy: 0.999 for English and 0.996 for Spanish. Averaged over these two languages, the mean accuracy is approximately 0.998.

For the ten AmericasNLI languages, the mean accuracy is 0.4331. The difference between high- and low-resource groups is therefore about 56.4 percentage points. That is, the *same* widely deployed model behaves almost like a reliable semantic reasoner for English and Spanish, and like a barely-above-chance classifier for Indigenous languages of the Americas.

We can make the high- vs. low-resource comparison a bit more explicit. Let $\mathcal{L}_{\text{high}}$ be the set of high-resource languages (in our case, English and Spanish from XNLI), and let $\mathcal{L}_{\text{low}}$ be the set of low-resource languages (the ten AmericasNLI

Group	Languages	Mean Acc.
High-resource (XNLI)	en, es	0.998
Low-resource (AmericasNLI)	10 langs	0.433

Table 3: Aggregate zero-shot accuracy of XLM-R_{large} on XNLI (English, Spanish) and AmericasNLI (10 Indigenous languages).

languages). For each language ℓ, let $\mathrm{Acc}(\ell)$ denote the test accuracy of f_θ on that language. We define group-wise mean accuracies as

$$\begin{aligned} \overline{\mathrm{Acc}}_{\mathrm{high}} &= \frac{1}{|\mathcal{L}_{\mathrm{high}}|} \sum_{\ell \in \mathcal{L}_{\mathrm{high}}} \mathrm{Acc}(\ell), \\ \overline{\mathrm{Acc}}_{\mathrm{low}} &= \frac{1}{|\mathcal{L}_{\mathrm{low}}|} \sum_{\ell \in \mathcal{L}_{\mathrm{low}}} \mathrm{Acc}(\ell). \end{aligned} \tag{4}$$

The performance gap we highlight can then be written as

$$\Delta_{\mathrm{Acc}} = \overline{\mathrm{Acc}}_{\mathrm{high}} - \overline{\mathrm{Acc}}_{\mathrm{low}}. \tag{5}$$

In our runs, $\overline{\mathrm{Acc}}_{\mathrm{high}} \approx 0.998$ and $\overline{\mathrm{Acc}}_{\mathrm{low}} \approx 0.433$, so $\Delta_{\mathrm{Acc}} \approx 0.564$.

We also plot this contrast directly in Figure 2. The visual gap between the two bars is a concrete reminder that current multilingual models are far from offering a fair distribution of semantic understanding across languages.

5.4 Qualitative Error Analysis

To understand model failures beyond aggregate metrics, we manually inspect a small sample of misclassified Quechua examples (`quy`) from the AmericasNLI test set (five cases, sampled from our evaluation outputs). A consistent pattern is over-predicting entailment under high lexical overlap, even when the hypothesis adds unsupported details (gold: neutral) or reverses polarity (gold: contradiction). We also observe confusion between contradiction and neutral when negation or evidential markers shift commitment in ways that are not captured by shallow lexical cues. Overall, these errors suggest the model often relies on surface overlap rather than tracking the underlying entailment relation, consistent with limited task-specific exposure to Indigenous morpho-syntax and discourse markers during fine-tuning.

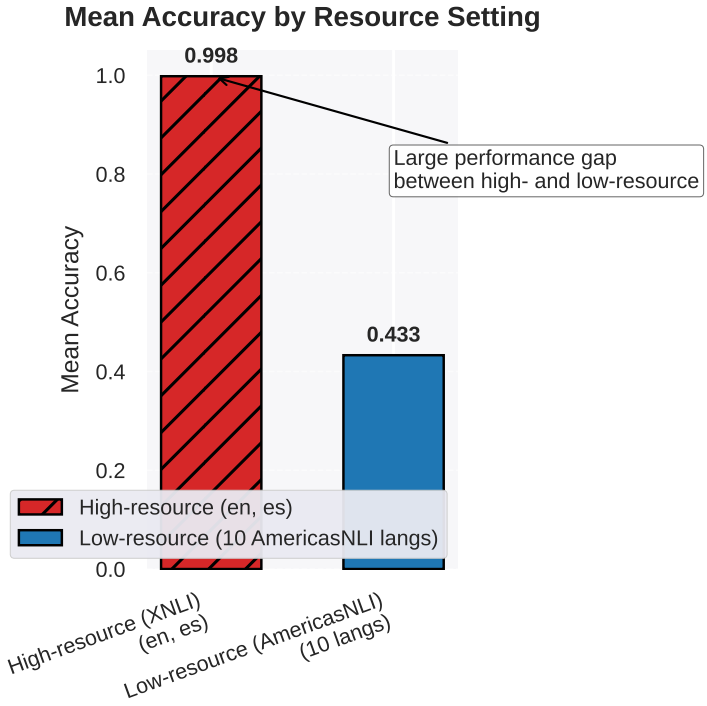

Fig. 2: Mean zero-shot accuracy of XLM-R_{large} for high-resource languages (English (en), Spanish (es)) vs. low-resource AmericasNLI languages.

6 Discussion

We now return to our three research questions and ask what the numbers mean for equity and fairness in multilingual NLP.

6.1 RQ1: How Well Does the Model Perform on Low-resource Languages?

On XNLI, XLM-R_{large} essentially solves the task: near-perfect accuracy and macro-F1 for English and Spanish. On AmericasNLI, by contrast, the average accuracy is only 43.3%, with all languages tightly clustered between roughly 40% and 47%.

From a purely machine learning perspective, one might say the model is "above chance" and move on. From a user-centered perspective, however, these numbers mean that for speakers of Indigenous American languages, a model that appears state-of-the-art in English behaves more like a rough heuristic than a dependable tool.

6.2 RQ2: Are There Systematic Performance Gaps Across Languages?

The group-level gap of around 56 percentage points between high- and low-resource languages is hard to ignore. It mirrors, in a concrete NLI setting, what prior work has reported at scale: language technologies work best for languages with large digital footprints and leave many other communities behind.

Within the AmericasNLI group, variation is relatively small. Shipibo-Konibo and Nahuatl sit at the high end, Rarámuri and Wixarika at the low end, but the spread is only about six percentage points. This suggests that the main issue is not fine-grained linguistic typology, but the fact that none of these languages were treated as first-class targets during model development.

6.3 RQ3: What Do Error Patterns Tell Us About Cultural and Linguistic Mismatch?

The Quechua examples show that the model struggles precisely where local morpho-syntax and pragmatics carry much of the meaning: negation, evidentiality, and subtle shifts in commitment or stance. It tends to over-predict entailment when there is high lexical overlap, even when the hypothesis is neutral or contradictory, and to over-predict contradiction when it encounters explicit negation without tracking what is actually being negated.

For speakers, this is more than an abstract optimization issue. Misclassifying contradictions as entailments effectively asserts that two statements agree when they in fact conflict. Collapsing neutral into contradiction can create an impression of disagreement where none exists. When such behaviours systematically affect already underserved communities, they become fairness concerns rather than mere model quirks.

6.4 Implications for Evaluation and Model Selection

Our results also speak to how we evaluate and select multilingual models. If we only look at averages over high-resource languages, or macro-averages over mostly well-supported languages, severe disparities can remain hidden. A model may be "state-of-the-art" on a popular benchmark and still perform at near-chance level for entire language families.

In this sense, AmericasNLI serves as a useful stress test: it exposes the limits of zero-shot transfer for languages that have, until recently, been almost invisible in large-scale NLP evaluation. For practitioners, the message is simple but important: whenever possible, we should inspect how a multilingual model behaves on truly low-resource languages, not just on those that dominate existing leaderboards.

7 Limitations and Future Work

This study is intentionally narrow. We focus on a single task (natural language inference), a single benchmark (AmericasNLI), and one widely used multilingual

model (XLM-R_{large} fine-tuned on XNLI). This makes the setup easy to reproduce and interpret, but also means our conclusions should be read with caution.

First, we do not adapt or fine-tune the model on Indigenous languages. Prior work shows that continued pretraining or translation-based approaches can substantially improve performance on AmericasNLI. Our goal here is different: to quantify out-of-the-box behavior in a realistic zero-shot setting. Future work should compare a range of adaptation strategies through the same fairness lens.

Second, we analyze only a small set of qualitative errors for one language (Quechua). A fuller picture would require systematic human analysis across all ten languages, ideally involving fluent speakers and community members rather than external annotators alone.

Third, our evaluation is purely text-based. Many real-world uses of NLP in Indigenous communities involve speech, code-switching, non-standard orthographies, or multimodal input. The gap we observe on clean written NLI data is therefore likely an optimistic estimate.

Finally, we do not attempt to model the social or historical forces that produced the current distribution of resources and benchmarks. Our work sits downstream of those structural inequalities; addressing them will require collaboration well beyond the technical NLP community.

8 Ethical Considerations

8.1 Working with Indigenous Languages

AmericasNLI was created with care, with documented data collection and annotation. Even so, work with Indigenous languages must attend to consent, representation, and benefit sharing. In this study we rely only on the publicly released splits and do not scrape additional data or train new models from scratch.

Our analysis also remains at arm's length from the communities whose languages are represented. We do not claim to speak for them or fully capture the social meaning of the sentences. Any deployment of models for these languages should involve collaboration with community members, local experts, and relevant institutions.

8.2 Risks of Misuse and Misinterpretation

A first risk is that poor performance is misread as evidence that Indigenous languages are "too hard" or "not worth supporting." The opposite is true: the gap we see reflects design choices and resource allocation in the NLP ecosystem, not properties of the languages themselves.

A second risk is over-confidence in models for high-resource languages. Near-perfect accuracy on XNLI for English and Spanish does not mean these models are unbiased or harmless. They can still encode and amplify stereotypes, treat social groups unfairly, or fail in unexpected ways outside their original training domains.

8.3 Fairness and Responsibility

From a fairness perspective, the core message is simple: current multilingual language models offer very different levels of service depending on which language a person speaks. This inequality is the outcome of choices about which languages to include in pretraining, which benchmarks to prioritize, and which adaptation methods to develop.

As researchers and practitioners, we have a responsibility to make these disparities visible and to push for evaluation practices that do not quietly ignore underserved communities. Benchmarks like AmericasNLI are one part of that work; building inclusive datasets, governance frameworks, and community partnerships is the longer-term task.

9 Conclusion

We set out to answer a simple question: what happens when a widely used multilingual NLI model, one that performs nearly perfectly on English and Spanish, is asked to handle truly low-resource Indigenous languages?

Using AmericasNLI as our testbed, we found that XLM-R_{large} in a zero-shot setting hovers just above chance for all ten Indigenous languages, with an average accuracy of 43.3%. In sharp contrast, the same model reaches almost 99.8% mean accuracy on XNLI for English and Spanish. The resulting gap of about 56 percentage points is a clear illustration of the linguistic digital divide.

Our qualitative analysis of Quechua error cases suggests that the model especially struggles with morpho-syntactic features and pragmatic nuances that were not part of its fine-tuning regime. In those regions of the space, it falls back on shallow lexical cues and produces label assignments that are often at odds with human intuition.

From a technical point of view, these results are unsurprising: it is hard to do well on languages that receive little or no explicit supervision. From a human point of view, however, they are a reminder that talk of "multilingual" models can easily obscure who actually benefits from the technology. If we care about NLP equity, we need evaluation setups that foreground underserved languages and make performance gaps impossible to ignore.

We hope that this small, focused study can serve as a compact, reproducible example for students and practitioners who want to engage seriously with fairness in multilingual NLP, even when working under tight time or resource constraints.

References

1. Bird, S.: Local languages, third spaces, and other high-resource scenarios. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7817–7829. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.539

2. Blasi, D.E., Anastasopoulos, A., Neubig, G.: Systematic inequalities in language technology performance across the world's languages. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5486–5505. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.376
3. Clark, J.H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., Palomaki, J.: TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics **8**, 454–470 (2020). https://doi.org/10.1162/tacl_a_00317
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL. pp. 8440–8451 (2020)
5. Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S.R., Schwenk, H., Stoyanov, V.: XNLI: Evaluating cross-lingual sentence representations. In: Proceedings of EMNLP. pp. 2475–2485 (2018)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
7. Ebrahimi, A., Mager, M., Oncevay, A., Chaudhary, V., Chiruzzo, L., Fan, A., Ortega, J., Ramos, R., Rios, A., Meza Ruiz, I.V., Giménez-Lugo, G., Mager, E., Neubig, G., Palmer, A., Coto-Solano, R., Vu, T., Kann, K.: AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6279–6299. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.435
8. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 4411–4421. PMLR (2020)
9. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6282–6293. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.560
10. Lauscher, A., Ravishankar, V., Vulić, I., Glavaš, G.: From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 4483–4499. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.363
11. Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S.O., Muhammad, S.H., Kabongo, S., Osei, S., et al.: Participatory research for low-resourced machine translation: A case study in African languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2144–2160. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.195
12. Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., Johnson, M.: XTREME-R: Towards more challenging and nuanced multilingual evaluation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 10215–10245. Association

for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.802
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30, pp. 5998–6008 (2017)