Evaluating the Diversity, Equity and Inclusion of NLP Technology: A Case Study for Indian Languages

Simran Khanuja[*]
Carnegie Mellon University[†]
skhanuja@andrew.cmu.edu

Sebastian Ruder[*]
Google Research
ruder@google.com

Partha Talukdar
Google Research
partha@google.com

Abstract

In order for NLP technology to be widely applicable, fair, and useful, it needs to serve a *diverse* set of speakers across the world's languages, be *equitable*, i.e., not unduly biased towards any particular language, and be *inclusive* of all users, particularly in low-resource settings where compute constraints are common. In this paper, we propose an evaluation paradigm that assesses NLP technologies across all three dimensions. While diversity and inclusion have received attention in recent literature, equity is currently unexplored. We propose to address this gap using the Gini coefficient, a well-established metric used for estimating societal wealth inequality. Using our paradigm, we highlight the distressed state of current technologies for Indian (IN) languages (a linguistically large and diverse set, with a varied speaker population), across all three dimensions. To improve upon these metrics, we demonstrate the importance of region-specific choices in model building and dataset creation, and more importantly, propose a novel, generalisable approach to optimal resource allocation during fine-tuning. Finally, we discuss steps to mitigate these biases and encourage the community to employ multi-faceted evaluation when building linguistically diverse and equitable technologies.

1 Introduction

NLP has seen large advances in recent years driven by the rapid progress in transfer learning (Ruder et al., 2019; Devlin et al., 2019). The benefits of these advances, however, are not equally distributed across the world's languages (Joshi et al., 2020) and users. While linguistic diversity and inclusion have evolved to be a pressing concern today, measures to quantify these are still lacking. The progress of any field is tightly coupled with its evaluation paradigm and the community is incentivized to work on highly visible metrics and benchmarks. In order for users around the world to reap the benefits of NLP technology, we must move from an evaluation that focuses on optimizing raw performance on available test data to a more holistic user-centric evaluation (Ethayarajh and Jurafsky, 2020; Ruder et al., 2021). In this paper, we attempt to do so by defining an evaluation framework along three dimensions: diversity, equity, and inclusion.[1]

Diversity is important as NLP technology should be available to speakers of any language (European Language Resources Association, 2019). To this end, recent work (Blasi et al., 2022) quantifies diversity of NLP technology across the world's languages by weighing normalized task performance for each language based on its speaker population.

Equity is key as we should aim to develop technology that does not discriminate against speakers of any particular language (Kaneko and Bollegala, 2019). State-of-the-art multilingual models in fact have been shown to perform much better in languages with access to many pre-training resources (Hu et al., 2020). To measure such performance inequity across languages, we propose to use the Gini coefficient (Dorfman, 1979), a measure that has been used to represent the income inequality within social groups.

Finally, inclusion is a concern as the fact that NLP technology is performant in a given task and language does not mean that it is usable by all. State-of-the-art models are becoming larger and larger (Fedus et al., 2021) and the low-resource settings of many languages often coincide with constraints on computational resources (Ahia et al., 2021). The value a technology provides to a user thus also needs to consider how easily such technol-

---

[*]Equal contribution.
[†]Work done at Google Research.

[1]We focus on assessing these dimensions on the *language level*. Prior work on equity focuses mainly on subpopulations *within* a language (Katell et al., 2020).

ogy can be deployed in practice. Ma et al. (2021) quantify this based on a model's runtime efficiency, considering factors like throughput and memory.

Our proposed paradigm is language and model agnostic making it applicable to an arbitrary set of languages and models. We apply our paradigm to highlight the distressing state of current technologies for Indian (IN) languages. India is a multilingual society with 1369 rationalized languages and dialects being spoken across the country (Chandramouli, 2011). Of these, 22 scheduled languages[2] spoken by almost 97% of the population hold an official recognition and 121 languages have more than 10,000 speakers. Additionally, 21.92% of its population lives below the poverty line (RBI, 2021). Serving this large varied population justly requires a multi-faceted effort and basing our case study on IN languages directs the way forward.

We evaluate a range of state-of-the-art models and transfer settings (Hu et al., 2020) across four standard downstream tasks: *Named Entity Recognition* (NER), *Part-of-Speech Tagging* (POS), *Natural Language Inference* (NLI) and *Question Answering* (QA). We observe that region-specific choices, i.e., a) region-specific pre-trained models (Kakwani et al., 2020; Khanuja et al., 2021) and b) Hindi as the transfer language during fine-tuning, generally yield the best results. In terms of efficiency, we find that smaller models are preferable for easier, syntactic tasks while larger models have the edge on more complex, semantic tasks.

Our findings, however, also highlight that we are still a long way from building perfectly inclusive and equitable NLP technology. Towards bridging this gap, we explore how we can most effectively annotate data for the remaining languages. Past work (Lin et al., 2019; Ahuja et al., 2022) has relied on heuristic and feature-based approaches to source language selection. In our work, we propose a novel, fully computational approach to model the space of source and target languages, and derive the optimal allocation of a fixed annotation budget to maximize performance on our proposed metrics in a multi-source setting.

Our contributions are the following: **1)** We propose a holistic evaluation paradigm that assesses NLP technology based on their diversity, equity,

and inclusion. **2)** Using this paradigm, we evaluate model capabilities for IN languages and quantify their shortcomings. **3)** We propose a novel approach to select data for fine-tuning these models with the objective of maximizing performance on the proposed metrics. **4)** We discuss steps that must be taken to mitigate these biases and call upon the community to incorporate our evaluation paradigm when building models to track progress towards building linguistically inclusive and diverse technologies.

2 Background and Related Work

Multilingual Models Transformer-based language models (LMs) (Vaswani et al., 2017) trained on massive amounts of text from multiple languages have enabled the inclusion of an unprecedented number of languages in NLP technologies (Conneau et al., 2020; Devlin et al., 2018). However, previous research has shown that these models do not serve all languages equally, with resource-poor languages in the long tail suffering the most (Hu et al., 2020; Lauscher et al., 2020). These models go through a critical step of fine-tuning for the downstream task before being deployed. Several recent works focus on optimal fine-tuning strategies that mitigate transfer gaps and improve overall performance across target languages. Lin et al. (2019) propose a tool that chooses optimal transfer languages based on linguistic features. Lauscher et al. (2020) demonstrate the effectiveness of investing in few-shot in-language training examples. Recently, Debnath et al. (2021) show that investing in an equal number of fine-tuning instances across target languages performs best. These past approaches however, have all been heuristically designed based on the knowledge and intuition of the experimenter.

User-centric Evaluation At its core, the need for language diversity in technologies is tied to the people it serves. Previous work (Ethayarajh and Jurafsky, 2020; Ma et al., 2021) has highlighted the need for transparent and user-centric leaderboard evaluation, reporting practically relevant statistics such as model size, energy efficiency, and inference latency. It is common for speaker populations of under-represented languages to operate in resource-constrained settings. Therefore, in addition to evaluating *linguistic* diversity, we follow Ma et al. (2021) in computing model efficiency, which serves to assess the *inclusivity* of these technologies. With regards to linguistic di-

---

[2]Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kashmiri, Kannada, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Tamil, Telugu, Sanskrit, Santali, Sindhi, Urdu

versity, Ruder et al. (2021) highlight the need for more fine-grained evaluation across languages and introduce language-specific leaderboards. Blasi et al. (2022) quantify the value of NLP technology weighed by speaker population and determine utilities of several technologies across the world's languages. Choudhury and Deshpande (2021) propose strategies for fair and efficient model selection depending on one's application, based on the principles of fairness in economics and social choice theory.

Indian Languages The research community has actively been contributing to the advancement of IN NLP by collecting and open-sourcing data (Kakwani et al., 2020; Ramesh et al., 2021; Abraham et al., 2020; Roark et al., 2020; Kunchukuttan et al., 2017; Khanuja et al., 2020a), building region-specific multilingual models (Khanuja et al., 2021; Kakwani et al., 2020; Ramesh et al., 2021) and creating evaluation benchmarks (Kakwani et al., 2020; Khanuja et al., 2020b)[3] Several of these efforts have been undertaken by AI4Bharat[4], a non-profit open-source community that has additionally been working on developing resources for IN signed languages (Sridhar et al., 2020) and creating keyboards for IN scripts. Recently, Google Research India launched a question answering (QA) challenge named ChAII[5]. Microsoft Research India has also made significant contributions to IN NLP with several efforts directed towards code-mixed language processing[6] and building tools and datasets for under-represented languages in India[7].

3 Diversity, Equity and Inclusion (DEI)

There is increasing awareness in society to promote diversity, equity and inclusion in our workforce, wherein such measures have recently been enforced by law (Constitution, 2021). In the social construct, *diversity* is defined as "the practice of including the many communities, identities, races, ethnicities, backgrounds, abilities, cultures, and beliefs of the people, including underserved communities", *equity* refers to "the consistent and systematic fair, just, and impartial treatment of all individuals" and *inclusion* means "the recognition, and use of the

talents and skills of employees of all backgrounds" (Constitution, 2021). Given the ubiquitous use of technology in our daily lives, we as technology makers hold the responsibility of making sure all voices are heard and equally represented in the technology we serve. Given that our research community is incentivized to work on highly visible metrics and benchmarks, an important first step is to encourage evaluation along these dimensions. Previous work mainly focused on average performance (as measured by accuracy or F1 for NLU tasks), which is not indicative of differences in DEI. Hence, while models claim state of the art based on an increase in average performance, this increase may only be due to making the "rich richer" (see Table 4).

We propose an evaluation paradigm for current NLP technology that operationalizes the well-established diversity, equity and inclusion pillars on a language level: we quantify diversity based on the value diverse speaker populations derive from a technology, equity based on egalitarian performance across speaker populations, and inclusion based on a technology's accessibility. We employ metrics of Blasi et al. (2022) and Ma et al. (2021) to measure diversity and inclusion respectively and propose a new metric to quantify equity. We describe the metrics in more detail below:

3.1 Diversity: *Utility, Demand and the Global Metric*

The global metric introduced by Blasi et al. (2022) helps quantify linguistic diversity. Formally, this metric is composed of the utility of a technology weighed by its demand. The utility u_l of a system for a task and language is its performance normalized by the best possible performance (typically, human-level performance) afforded by the task:

$$u_l = \frac{\text{performance}_l}{\text{theoretical max performance}}$$

Demand d_l is characterized by taking into consideration demographic and linguistic perspectives. Under the demographic perspective, the demand for a given technology in a language is estimated to be proportional to the number of speakers of the language itself n_l ($d_l \propto n_l$). Under the linguistic perspective, the demand across languages is identical ($d_l \propto 1$). These two alternatives, as well as any intermediate combination of them, are parameter-
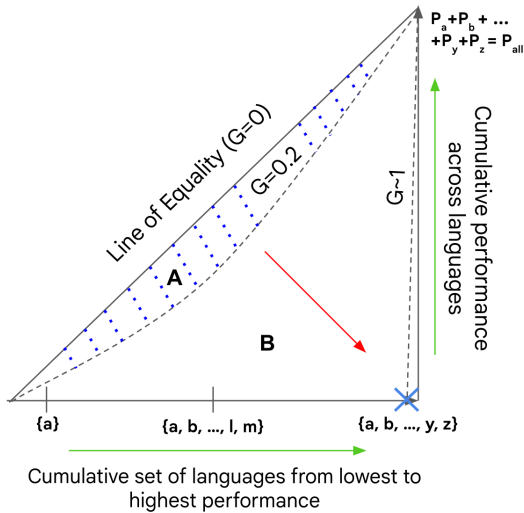

Figure 1: *Graphical Representation* of the Gini coefficient (G), given by $A/(A + B)$ when G = 0.2. Each point on the graph depicts the proportion of the total (cumulative) performance P_{all} (e.g., accuracy, F1, etc) that is achieved by the bottom $n\%$ of languages combined. Assume we have a language set {a, b, ..., y, z} with performances $P_a \leq P_b ... \leq P_y \leq P_z$ and $P_a + P_b ... P_y + P_z = P_{all}$. When all languages perform the same, i.e., $P_a = P_b = ... = P_y = P_z$, G = 0, as represented by the line of equality, i.e., the bottom $n\%$ of languages also account for $n\%$ of the total performance. The value of G increases as the disparity in performance between all languages increases, and approaches unity in the case of perfect inequality (here, this would mean $P_a = P_b = ... = P_y = 0$ and $P_z = P_{all}$), i.e., the model / application supports only one language. See §3.2 for details.

ized through a single exponent τ:

$$d_l^{(\tau)} = \frac{n_l^\tau}{\sum_{l' \epsilon L} n_{l'}^\tau}$$

where $\tau = 1$ corresponds to a demographic notion of demand and $\tau = 0$ to a linguistic one. The global metric can now be defined as:

$$M_\tau = \sum_{l \epsilon L} d_l^{(\tau)} \cdot u_l$$

In essence, $M_\tau = 0$ means that no user benefits from language technology and $M_\tau = 1$ corresponds to each language user enjoying perfect technology. Given our people-centric aim to measure benefit for all speakers, we employ the demographic notion of demand ($M_{\tau=1}$).

3.2 Equity: *Gini Coefficient*

While diversity accounts for a language's speaker population, it does not take into account inequal-

ities in the performance across languages. While several past works have highlighted transfer gaps in performance across languages (Hu et al., 2020), none have quantified this dispersion.[8] Traditionally used measures of statistical dispersion like standard deviation or calculating range are sub-optimal choices as they are scale-dependant, unbounded and highly sensitive to outliers, which makes them unsuitable for data that does not approach a normal distribution (De Maio, 2007).

Beyond these measures, several nuanced metrics have been introduced to quantify disparity in income distributions. The choice of income inequality indicator is of significant importance since it has implications in measuring health, state-level mortality, etc. (De Maio, 2007). The Gini coefficient (Dorfman, 1979) has been most commonly used for this purpose (De Maio, 2007).

Hurley and Rickard (2009) lay out six desirable attributes of a measure of sparsity, drawing from past literature (Dalton, 1920; Rickard and Fallon, 2004) and prove the Gini coefficient to be the only measure having all six, among a varied set of alternatives. Briefly, these properties and their relevance in measuring linguistic disparity across tasks include: i) *Robin Hood*: a drop in high-performant and gain in low-performant languages should lead to higher equity; ii) *Scale Invariance*: no change in relative performance should lead to no change in equity, regardless of changes in absolute values; iii) *Rising Tide*: adding a constant value to each language's performance should increase equity; iv) *Cloning*: equity must remain invariant under cloning, i.e., if two identical distributions are combined, the equity remains unchanged; v) *Bill Gates*: if one language hypothetically gains infinite performance, equity should tend to zero; vi) *Babies*: adding languages with zero performance in the distribution should decrease equity.

Given that the Gini coefficient satisfies all of these attributes (Hurley and Rickard, 2009), we propose to use the same in pursuit of quantifying the inequalities amongst languages with regard to downstream tasks in NLP. A pictorial representation of the Gini coefficient for this setting can be found in Figure 1. Downstream task performance closely follows the highly skewed data distributions on which massively multilingual models are pre-trained. By including the Gini coefficient mea-

---

[8]Hu et al. (2020) only considered the difference between English and other languages as cross-lingual transfer gap.

sure in our evaluation, we aim to incentivize model builders to invest in equitable performance, despite data differences. Details on how the Gini coefficient is calculated are given in Appendix A.3.

3.3 Inclusion: *Efficiency Score*

Language technology is only beneficial if it can be deployed and accessed by users in a region. We employ efficiency to quantify inclusion as user devices are resource-constrained in many low-resource settings. Following work on user-centric evaluation (Ethayarajh and Jurafsky, 2020; Ma et al., 2021), we propose to incorporate efficiency into model performance based on throughput and memory, each of which are defined below.[9]

Throughput Number of instances the model can process per second on a CPU, assuming that GPUs are rarely used for deployment at scale in resource-constrained environments.

Memory Saved The size of the model is considered to be a measure of how expensive a model is to use in practice. Since we wish to minimize this metric, *memory used* is transformed into *memory saved* by subtracting it from a maximum available memory of 16 GB (Ma et al., 2021). We show the memory and throughput values for our models in Appendix A.1.

Following Ma et al. (2021), to calculate the efficiency score we first convert each metric into units of performance, by calculating the average marginal rate of substitution (AMRS) for each metric M (i.e., throughput and memory). $\mathrm{AMRS}(M, \mathrm{perf})$ tells us the rate at which model creators, as a group, are trading off M for a one-point increase in perf while keeping utility constant. For example, if AMRS of "memory saved" with respect to accuracy were 0.5 GB, then each GB of memory saved would on average be worth 2 points of accuracy. Dividing M by $\mathrm{AMRS}(M, \mathrm{perf})$ converts it to units of performance. Details on how one can calculate $\mathrm{AMRS}(M, \mathrm{perf})$ can be found in Appendix A.1. For a model x_i, $\mathrm{Efficiency}(x_i)$ is then defined as :

$$\mathrm{Efficiency}(x_i) = \sum_M w_M * \frac{M(x_i)}{\mathrm{AMRS}(M, \mathrm{perf})}$$

where we choose $w_{\mathrm{perf}} = 0.5$, $w_{\mathrm{throughput}} = 0.25$ and $w_{\mathrm{memory}} = 0.25$ as default weights. In practice, these weights can be adjusted based on user requirements and existing constraints.

4 Egalitarian Annotation Budget Allocation

Model development involves not just the design of an architecture or training but also data annotation. The proposed dimensions thus cannot only be used to assess models but can also inform how data should be annotated across many languages. As fine-tuning on a few labeled examples in the target language has shown to improve zero-shot transfer performance, we study how to allocate an annotation budget across a number of source languages S in order to optimize for inclusion and equity across a set of target languages T. Previous work employs a feature-based approach to select a single source language to maximize performance on a target language (Lin et al., 2019) or labels examples across all source languages equally (Debnath et al., 2021). We propose a fully computational approach for modeling the space of source and target languages for a multi-source multi-target language setting. This is done by empirically estimating performance of language $t \, \epsilon \, T$ on a held-out set when fine-tuned on x labeled instances of language $s \, \epsilon \, S$, $\forall (s, t)$ pairs, which follows a power-law distribution (Rosenfeld et al., 2019). We now seek to find the optimal allocation $\{x_s : s \in S\}$ subject to $\sum_{s \in S} x_s \leq X$ (details in Appendix A.5).

We follow a simple greedy approach to solve this constrained optimization problem as shown in Table 10. Specifically, at each step we allocate a sample to the source language conferring the highest marginal gain to all target languages, which is quantified by the summation of the increase in the global metric and the reduction in Gini.[10] At present, we assign equal weight to each metric but this can be changed according to user preferences.

5 Experiments

5.1 Experimental setup

Languages We base our case study on the 22 scheduled languages of India spoken by 97% of its population. We also include English, since it has a sizeable population of 128.5M speakers (Table 1).

---

[9]Ma et al. (2021) additionally consider fairness and robustness, both of which are highly contextual and difficult to define in the context of multilingual models at present. Hence, we focus on model aspects that are objectively measurable.

[10]Future work may consider more complex approaches that consider language relatedness based on work on transfer relationship learning (Zamir et al., 2018; Song et al., 2019).

Language	as	bn	brx	doi	en	gu	hi	kn	kok	ks	mai	ml
Speakers (in M)	23.6	107.4	1.6	2.8	128.5	60.3	691.6	58.8	2.6	7	14.3	35.6
Language	mni	mr	ne	or	pa	sa	sat	sd	ta	te	ur	-
Speakers (in M)	2.2	99.1	3.4	42.6	36.1	3.1	7.7	3.1	76.6	94.5	63.2	-

Table 1: The number of speakers (in millions) for each of the 22 scheduled languages and English. We take the sum total of first, second and third language speakers for each language.

Task	Dataset	Test Langs.	HP
NER	WikiAnn (Pan et al., 2017; Rahimi et al., 2019)	bn, en, gu, hi, ml, mr, pa, ta, te, ur	97.6
POS	Universal Dependencies v2.6 (Nivre et al., 2018)	en, hi, mr, ta, te, ur	97
NLI	XNLI (Conneau et al., 2018)	en, hi, ur	92.8
QA	XQuAD (Artetxe et al., 2019); TyDiQA-GoldP (Clark et al., 2020)	bn, en, hi, te	91.2; 90.1

Table 2: *Finetuning Tasks and Datasets.* HP denotes the human performance for each task. For QA, HP is 91.2 F1 for XQuAD and 90.1 F1 for TyDiQA.

Language	NER	POS	NLI	QA
English	20,000	21,261	392,702	88,602
Hindi	5,000	13,305	392,702 (*-tran*)	88,602 (*-tran*)

Table 3: *Number of training instances for English and Hindi.* (*-tran*) denotes that the English fine-tuning set has been translated to Hindi.

Task	Model	Baseline $F1/Accuracy \uparrow$	Diversity $M_{\tau=1} \uparrow$	Equity $Gini\ Coeff. \downarrow$	Inclusion $Efficiency \uparrow$
NER	MuRIL$_{base}$	77.6	69.6	0.59	69.1
	XLM-R$_{large}$	68.0	61.2	0.60	44.4
	MuRIL$_{large}$	77.7	68.2	0.59	63.1
POS	MuRIL$_{base}$	75.0	54.7	0.76	52.5
	XLM-R$_{large}$	79.2	60.3	0.75	48.0
	MuRIL$_{large}$	77.3	58.6	0.76	51.8
NLI	MuRIL$_{base}$	74.1	45.5	0.88	58.7
	XLM-R$_{large}$	78.7	46.6	0.88	57.3
	MuRIL$_{large}$	78.6	47.4	0.88	57.8
QA	MuRIL$_{base}$	76.1	53.8	0.83	77.8
	XLM-R$_{large}$	75.7	56.6	0.83	76.3
	MuRIL$_{large}$	77.7	57.9	0.83	75.7

Table 4: *DEI Results compared to baseline F1/accuracy performance.* Here, we compare models' accuracy/F1 performances (usually reported as the evaluation metric) to their DEI metrics. We observe that while performances may significantly vary, DEI metrics (especially equity) don't change as much, indicating that multilingual models make the rich "richer" to increase average performance but may not be moving towards being truly multilingual and equitable across languages. More discussions in Section 5.2.

Tasks We select tasks from the XTREME (Hu et al., 2020) benchmark. Dataset details and the human performance (HP) for each task can be found in Table 2. For each task, we only evaluate on IN language test sets.

Models Model selection is motivated by two key factors that we wish to explore in our study: **i)** general vs region-specific choices; and **ii)** model efficiency. We choose IndicBERT (Kakwani et al., 2020), MuRIL (Khanuja et al., 2021) and XLM-R (Conneau et al., 2020), the first two being region-specific models and the third being a state-of-the-art model trained on 100+ languages. We consider both the base and large versions for MuRIL and XLM-R. IndicBERT follows the ALBERT architecture (Lan et al., 2019) and is hence much smaller than the base versions of both models. IndicBERT is trained on 11, MuRIL on 16, and XLM-R on 15 IN languages (details in Appendix A.2).

Fine-tuning We initially fine-tune the selected models using training data in English (EN) given the availability of labeled data across tasks. However, past works highlight that this choice is sub-optimal and one can obtain better performance by transferring from closely related languages (Lauscher et al., 2020; Cotterell and Heigold, 2017; Dong et al., 2015; Turc et al., 2021). To examine this effect in our case study, we additionally fine-tune models on Hindi (HI) because **i)** 15 out of 22 languages belong to the same language family as HI (Indo-Aryan); **ii)** we have training data available for all tasks in HI[11]; and **iii)** HI has the highest speaker population, which may lead to higher demographic utility and is a future-safe choice to obtain annotations for any task. Table 3 summarizes training data statistics for EN and HI.

---

[11] Training sets for NLI and QA have been machine-translated from English, which has been shown to perform similar to human-generated train sets (Turc et al., 2021).

Metric	Train Lang.	Model	NER	POS	NLI	QA	Average
$M_{\tau=1}$ ↑ (Diversity)	English	MuRIL$_{base}$	69.6	54.7	45.5	53.8	55.9
		XLM-R$_{large}$	61.2	60.3	46.6	56.6	56.2
		MuRIL$_{large}$	68.2	58.6	47.4	**57.9**	58.0
	Hindi	MuRIL$_{base}$	**75.1**	**67.3**	46.8	54.7	61.0
		XLM-R$_{large}$	74.4	66.8	**49.4**	53.2	60.9
		MuRIL$_{large}$	74.8	66.5	49.2	54.6	**61.3**
Gini Coeff. ↓ (Equity)	English	MuRIL$_{base}$	0.59	0.76	0.88	0.83	**0.76**
		XLM-R$_{large}$	0.6	0.75	0.88	0.83	0.77
		MuRIL$_{large}$	0.59	0.76	0.88	0.83	0.77
	Hindi	MuRIL$_{base}$	0.59	0.75	0.87	0.83	**0.76**
		XLM-R$_{large}$	0.59	0.76	0.88	0.83	0.77
		MuRIL$_{large}$	0.59	0.75	0.87	0.83	**0.76**
Efficiency ↑ (Inclusion)	English	MuRIL$_{base}$	69.1	52.5	58.7	77.8	64.5
		XLM-R$_{large}$	44.4	48	57.3	76.3	56.5
		MuRIL$_{large}$	63.1	51.8	57.8	75.7	62.1
	Hindi	MuRIL$_{base}$	**69.8**	**56.2**	**59.8**	77.3	**65.8**
		XLM-R$_{large}$	49.2	49.0	59.1	75.1	58.1
		MuRIL$_{large}$	65.2	53.7	58.8	75.0	63.2

Table 5: *Region-specific fine-tuning results.* Note that the metrics are computed considering all 23 languages as detailed in Section 5.1. Region-specific fine-tuning helps, but disparities along DEI axes persist. More discussions in Section 5.2.

5.2 Zero-shot transfer results

How do DEI metrics compare to baseline standard performance metrics (F1/Accuracy)? We report results of the best-performing models in Table 4. While average performance is similar across tasks, there are stark differences in DEI metrics. The diversity metric helps discern whether the change in performance is more skewed towards languages with a relatively high or low speaker population. For example, for POS, MuRIL$_{base}$ and XLMR$_{large}$ have a 4.2% difference in performance but a 5.9% difference in $M_{\tau=1}$. This indicates that the difference is more pronounced for languages with large speaker populations. Similarly, for NLI, the difference in performance and $M_{\tau=1}$ is 4.6% and 1.1% respectively, which also highlights a lack of test data to quantify larger differences in diversity. With regards to equity, we observe that even though major differences exist compared to average performance, the Gini coefficient remains relatively unchanged, indicating that while overall performance has increased, the disparity in performance amongst languages has not yet been addressed by any model. Regarding efficiency or inclusion, while MuRIL$_{large}$ beats MuRIL$_{base}$ in performance, MuRIL$_{base}$ is more efficient to use, across all tasks.

Where are we today w.r.t DEI of NLP technology? We report results of best-performing models (fine-tuned on EN and HI) in Table 5 (detailed results with XLMR$_{base}$ and IndicBERT in Table 13). Overall, the diversity metric is highest

Metric	Budget	Model	Fine-tuning Strategy			
			English	Hindi	Egalitarian	Greedy
$M_{\tau=1}$ ↑	1,000	XLM-R$_{large}$	54.0	66.2	65.4	65.3
		MuRIL$_{large}$	60.4	71.3	**74.1**	73.6
	5,000	XLM-R$_{large}$	59.4	74.4	75.4	75.7
		MuRIL$_{large}$	65.4	74.8	78.2	**78.3**
	10,000	XLM-R$_{large}$	59.0	-	77.6	77.6
		MuRIL$_{large}$	70.5	-	79.6	**79.9**
Gini Coeff. ↓	1,000	XLM-R$_{large}$	0.6	0.6	0.59	0.59
		MuRIL$_{large}$	0.6	0.6	**0.58**	**0.58**
	5,000	XLM-R$_{large}$	0.6	0.59	0.59	0.59
		MuRIL$_{large}$	0.59	0.59	**0.58**	**0.58**
	10,000	XLM-R$_{large}$	0.61	-	0.59	0.59
		MuRIL$_{large}$	0.59	-	**0.58**	**0.58**

Table 6: *Performance on NER under different annotation budgets.* We observe that the greedy approach (§4) performs best across all metrics. Note that the HI train set has 5,000 examples only. Details in §5.3.

for MuRIL$_{large}$, when fine-tuned on HI. We also observe that the diversity metric increases with region-specific choices, both in pre-training and fine-tuning. The Gini coefficient remains relatively high at around 0.76 even for the best models, which highlights the disparity in performance even among languages within a single region.[12] With regards to efficiency, averaging across languages and tasks, MuRIL$_{base}$ performs best.

What is the way forward? Overall, the absolute values of the global metric and the Gini coefficient indicate that there lies great potential in both increasing the utility of our models and making them more equitable. Since model performances partially reflect the amount of raw data used in pre-training (Lauscher et al., 2020), creating equitable unlabeled data resources would alleviate these issues. However, this is an ambitious undertaking that is extremely resource intensive and can certainly not be achieved for 6500 languages in the near future. We thus investigate how limited amounts of data can be used to maximally improve utility and equity during fine-tuning.

5.3 Few-shot results

Problem Formulation For few-shot fine-tuning, we focus on NER where sufficient labeled training data for seven IN languages is available. We employ the source languages $S = \{bn, en, hi, ml, mr, ta, ur\}$ and seek to optimize metrics on the target languages $T = \{bn, en, gu, hi, ml, mr, pa, ta, te, ur\}$. In

---

[12]For comparison, for OECD countries from 2008–2009, the Gini coefficient on income for the entire population ranged between 0.34 and 0.53 while the Gini coefficient for the entire world has been estimated to be between 0.61 and 0.68 (Hillebrand et al., 2009; Klugman and Nations, 2010).

each setting, we have a limited annotation budget, which we can divide among the source languages. We compare against several competitive baselines: **i)** using only examples from EN or HI respectively; **ii)** distributing the annotation budget in an egalitarian (uniform) way across all source languages (Debnath et al., 2021); and **iii)** our novel greedy approach proposed in §4. For the greedy approach, we illustrate the best-fit curves for each (s, t) pair in Appendix A.5 (Table 11).

Results We show the results under various annotation budgets in Table 6. Overall, we find that our method yields a higher global metric under most budgets (5 of 6 cases) and also yields a lower Gini coefficient under *all* budget schemes. The optimal allocations for each budget are shown in Table 12. As we can see, the greedy algorithm converges to a solution that is close to uniform. This provides further evidence for the benefits of an egalitarian distribution of annotation budget in order to maximize performance across all languages as the expected marginal gain for languages that have been under-represented during training will be highest. Both the egalitarian and greedy approaches significantly outperform fine-tuning on EN or HI. For instance, our greedy approach outperforms fine-tuning on 10,000 EN examples by 1–3% with a budget of only 1,000 examples.

6 Discussion

Building evaluation datasets Having uncovered the linguistic inequity and exclusivity of current NLP technologies, we seek to identify practical measures we can take in order to mitigate these biases. As a first step, it is paramount to build representative evaluation sets for all languages as they are required to accurately measure diversity and equity. Out of the 23 languages in our case study, most do not have evaluation data across tasks despite holding official recognition and being spoken by 97% of the population. In light of the benefits of an egalitarian data distribution during few-shot learning, we also recommend the collection of small amounts of data across many languages for training, in order to maximize marginal gain. These datasets should be collected at the grass-roots level, involving the community they need to serve to capture culturally relevant phenomenon. A prime example of this is the Masakhane organisation[13] steering efforts towards data collection in

---

[13] https://www.masakhane.io/

African languages, involving the local community. Incentivizing rural, low-income workers to provide for such data also serves as a viable source of supplementary income, and does not degrade dataset quality (Abraham et al., 2020).

Trading off multilinguality and regionality From a modeling perspective, multilingual pre-trained models have been instrumental to NLP systems supporting an unprecedented number of languages, because of their zero-shot transfer capabilities. However, while these are a big step towards linguistic inclusion, they are subject to limitations such as highly skewed pre-training distributions and limited transfer to under-represented languages (Hu et al., 2020; Lauscher et al., 2020), a bias towards the source language, and sub-optimal tokenization (Wang et al., 2021). A way to combat these issues is to make region-specific choices, both in pre-training and fine-tuning, as observed in §5.2. Localizing the problem also enables one to incorporate linguistic expertise (Nzeyimana and Niyongabo Rubungo, 2022) and provide support for culturally relevant phenomena like transliteration or code-mixing. Despite this, we must be wary of excessive fragmentation in pre-training as it leads to higher maintenance costs and there is a possibility that these benefits will be overcome with advances in compute and model capacity in the near future. Optimal fine-tuning however, is promising, as evidenced in §5.3 where we observe significant gains in moving away from the zero-shot paradigm.

7 Conclusion

We have proposed a framework for the evaluation of NLP technology based on diversity, equity, and inclusion and proposed the Gini coefficient to quantify equity. We have assessed to what extent several modeling and data choices affect the value NLP technology confers to speakers of Indian languages. We have also proposed an algorithmic method for resource allocation for task-specific fine-tuning, which outperforms a purely egalitarian distribution of data labeling. Finally, we highlight the importance of building representative evaluation sets from the grass-roots level to enable tracking progress, and discuss how even with the best modeling strategies, we have a long road ahead in building inclusive, equitable systems. While region-specific choices help to a certain extent, building a single global multilingual model without compro-

mising on the three metrics is something we should move towards in the future. We sincerely hope our evaluation paradigm aids in tracking the community's progress in building linguistically diverse technologies.

Limitations

We do not consider the inequalities that may exist within subgroups in a language given the lack of fine-grained evaluation data. In multilingual countries like India, each language is composed of several dialects (Hindi alone is composed of 58 dialects (Chandramouli, 2011)). As disparities exist along multiple axes such as caste, gender, religion and so on (Sambasivan et al., 2021), it is imperative to go beyond the language level. We only consider pre-trained language models for our experiments given their massive language coverage and zero-shot transfer capabilities. There have been efforts to build language-specific, task-specific models which we do not include in our study. Our greedy data allocation method is a strong baseline that outperforms standard approaches such as selecting a single source language or uniform selection. It can be improved by incorporating notions of language similarity, which requires more complex methods (Song et al., 2019).

Acknowledgements

References

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2819–2826.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation. In *Findings of EMNLP 2021*.

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. 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)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Chandramouli. 2011. Census of india 2011 provisional population totals. *New Delhi: Office of the Registrar General and Census Commissioner*.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12710–12718.

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

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.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

US Constitution. 2021. US constitution, 2021. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/06/25/executive-order-on-diversity-equity-inclusion-and-accessibility-in-the-federal-workforce/.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual, character-level neural morphological tagging. *arXiv preprint arXiv:1708.09157*.

Hugh Dalton. 1920. The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361.

Fernando G De Maio. 2007. Income inequality measures. *Journal of Epidemiology & Community Health*, 61(10):849–852.

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, and Antonios Anastasopoulos. 2021. Towards more equitable question answering systems: How much more data do you need? *arXiv preprint arXiv:2105.14115*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732.

Robert Dorfman. 1979. A formula for the gini coefficient. *The review of economics and statistics*, pages 146–149.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of EMNLP 2020*.

European Language Resources Association. 2019. BLT4All: Language Technologies for All. https://lt4all.elra.info/en/. [Online; accessed Dec. 2019.].

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv preprint arXiv:2101.03961*.

Evan Hillebrand et al. 2009. Poverty, growth and inequality over the next 50 years. In *Expert Meeting on How to feed the World in*, volume 2050, pages 2012–02.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. In *Proceedings of ICML 2020*.

Niall Hurley and Scott Rickard. 2009. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of ACL 2020*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Binz, Daniella Raz, and P. M. Krafft. 2020. Toward situated interventions for algorithmic equity: Lessons from the field. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 45–55.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020a. A new dataset for natural language inference from code-mixed conversations. *arXiv preprint arXiv:2004.05051*.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020b. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*.

Jeni Klugman and Development Programme United Nations. 2010. *The real wealth of nations: pathways to human development*. Palgrave Macmillan.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of ACL 2019*.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *arXiv preprint arXiv:2106.06052*.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, and Antonsen et al. 2018. Universal Dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596*.

Reserve Bank of India RBI. 2021. Handbook of statistics on indian economy. *RBI, Reserve Bank of India*.

Scott Rickard and Maurice Fallon. 2004. The gini index of speech. In *Proceedings of the 38th Conference on Information Science and Systems (CISS'04)*.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Işın Demirşahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 2413–2423.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2019. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. In *Proceedings of EMNLP 2021*.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328.

Jie Song, Yixin Chen, Xinchao Wang, Chengchao Shen, and Mingli Song. 2019. Deep model transferability from attribution maps. In *Advances in Neural Information Processing Systems*.

Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. Include: A large scale dataset for indian sign language recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1366–1375.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view Subword Regularization. In *Proceedings of NAACL 2021*.

Amir R Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling Task Transfer Learning. In *Proceedings of CVPR 2018*.

A Appendix

A.1 Efficiency

We report the throughput and memory for each model and task in Table 7. For NLI, POS and NER, the maximum sequence length is 128 and for QA it's 384.

As detailed in Section 3.3, we need to calculate the AMRS for each metric M (throughput and memory saved), to calculate the efficiency score. As described in (Ma et al., 2021), each model has properties (or *goods*) that inform its utility. Here, these goods are throughput, memory saved, and performance. A model is a point in this space of goods and an indifference curve is a set of points

that provide the same utility (for different values of these properties). These curves are monotonically negatively sloped, i.e., for a model with higher accuracy to be on the same curve as one with a lower accuracy, it will have to use up more memory or have lower throughput. For a given indifference curve, the rate at which this trade-off is made, is called the marginal rate of substitution (MRS).

To calculate MRS, and consequently AMRS, (Ma et al., 2021) make two key assumptions: **i)** All models lie on the same indifference curve; **ii)** if $M(x_i) > M(x_{i+1})$ and $perf(x_i) > perf(x_{i+1})$, then there exists a model $\langle perf(x_{i+1}), M(x_i) + (M(x_i) - M(x_{i+1}))\rangle$ on the same indifference curve as x_i. For our case study, we believe that assuming regional and global models to lie on the same indifference curve would be inaccurate, since models with the same capacity (size and architecture) have been trained on a different set of languages. In the case of (Ma et al., 2021), they only consider models pre-trained on English. Here, we assume that regional models (trained on 15-17 languages) would be strictly better on all dimensions and hence lie on a different indifference curve as compared to global models (trained on 100+ languages). Hence, we assume IndicBERT, $MuRIL_{base}$ and $MuRIL_{large}$ to lie on one indifference curve and $XLMR_{base}$ and $XLMR_{large}$ to lie on another. The second assumption holds in our case as well.

For a model x_i, Efficiency(x_i), MRS, and AMRS are given by :

$$Efficiency(x_i) = \sum_M w_M * \frac{M(x_i)}{AMRS(M, perf)}$$

$$AMRS(M, perf) = \overline{MRS}$$

$$MRS = \left\{ \left| \frac{M(x_i) - M(x_{i+1})}{perf(x_i) - perf(x_{i+1})} \right| 1 \le i < n \right\}$$

A.2 Pre-training Languages

In Section 5.1, we choose IndicBERT, MuRIL and XLM-R as pre-trained multilingual models to base our analysis upon. IndicBERT is trained on 11 IN languages that include Assamese (as), Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Oriya (or), Punjabi (pa), Tamil (ta), Telugu (te). XLM-R includes 15 IN languages in training with the addition of Nepali (ne), Sanskrit (sa), Sindhi (sd) and Urdu (ur) over IndicBERT and MuRIL is trained on 16 IN languages, with the addition of Kashmiri (ks) over XLM-R.

Model	Metric	NER	PoS	NLI	QA
IndicBERT	Memory Saved	15.9GB			
	Throughput	22.6	20.2	22.9	10.5
	Perf (EN)	41.3	71.6	69.7	52.4
XLM-R$_{base}$	Memory Saved	15GB			
	Throughput	24.4	23.2	26.4	14.9
	Perf (EN)	61.7	82.2	77.1	72.1
MuRIL$_{base}$	Memory Saved	15.1GB			
	Throughput	23.8	23.1	26.2	15.7
	Perf (EN)	74.9	80.3	78.9	77.5
XLM-R$_{large}$	Memory Saved	13.9GB			
	Throughput	9.4	10.0	10.4	4.1
	Perf (EN)	64.6	83.7	81.8	81.4
MuRIL$_{large}$	Memory Saved	14.1GB			
	Throughput	9.8	9.9	10.5	4.2
	Perf (EN)	71.8	83.4	82.8	83.0
AMRS (Regional)	Throughput	1.7	3.4	2.2	1.1
	Memory	0.1	0.3	0.2	0.1
AMRS (Global)	Throughput	3.7	7.1	3.3	1.2
	Memory	0.3	0.5	0.2	0.1

Table 7: The throughput is given by the number of instances processed per second by the fine-tuned models on CPU.

A.3 Gini Coefficient

The Gini coefficient is mathematically computed based on the Lorenz curve, which plots the relation between population size and the cumulative income earned by that population as shown in Figure 1. To plot the Lorenz curve, individuals are sorted in increasing order of income (x-axis) and their cumulative wealth is plotted on the y-axis. In essence, a point (x, y) indicates that the bottom x% of the population holds y amount of wealth. The line at 45 degrees represents perfect equality of incomes. The Gini coefficient G is then calculated as the ratio of the area that lies between the line of equality and the Lorenz curve (A in Figure 1), over the total area under the line of equality (A + B in Figure 1). If G = 0, every person in the population receives an equal percentage of income and if G = 1, a single person receives 100% of the income. Since the axes scale from 0 to 1, A + B = 0.5. In essence, if the Lorenz curve is represented by the function Y = L(X) then G can be given as:

$$G = \frac{A}{A + B} = 2A = 1 - 2B = 1 - 2\int_0^1 L(X)dX$$

For a population with values y_i, i = 1 ... n, that are indexed in non-decreasing order ($y_i \le y_{i+1}$):

$$G = \frac{1}{n}\left(n + 1 - 2\frac{\sum_{i=1}^n (n + 1 - i)y_i}{\sum_{i=1}^n y_i}\right)$$

For comparison, for OECD countries from 2008–2009, the Gini coefficient on income for the entire

Metric	Train Lang.	Model	NER	PoS	NLI	QA	Average
Gini Coeff. ↓	English	IndicBERT	0.155	0.107	0.051	0.091	0.101
		XLM-R$_{base}$	0.095	0.067	0.058	0.048	0.067
		MuRIL$_{base}$	0.047	0.086	0.048	0.03	**0.052**
		XLM-R$_{large}$	0.084	0.06	0.049	0.026	0.055
		MuRIL$_{large}$	0.051	0.086	0.051	0.027	0.057
	Hindi	IndicBERT	0.173	0.073	0.004	0.041	0.073
		XLM-R$_{base}$	0.067	0.037	0.039	0.046	0.047
		MuRIL$_{base}$	0.062	0.032	0.036	0.012	**0.035**
		XLM-R$_{large}$	0.057	0.04	0.033	0.029	0.04
		MuRIL$_{large}$	0.065	0.057	0.033	0.014	0.042

Table 8: *Gini Coefficient* for all models calculated only across languages having evaluation sets for each task.

Task	Batch Size	Learning Rate	No. of Epochs	Warmup Ratio	Max. seq. Length
NER	32	2e-5	10	0.1	128
POS	32	2e-5	10	0.1	128
NLI	64	2e-5	3	0.1	128
QA	32	3e-5	2	0.1	384

Table 9: Hyperparameter details for each fine-tuning task

population ranged between 0.34 and 0.53. The Gini coefficient on income for the entire world has been estimated to be between 0.61 and 0.68 (Hillebrand et al., 2009; Klugman and Nations, 2010). In our experiments on Indian languages, state-of-the-art models achieve an average Gini coefficient of 0.77, which highlights the disparity in performance even among languages within a single region.

As mentioned in Section 5.2, calculating the Gini coefficient across all 23 languages doesn't reflect the dispersion in performances across languages for which we have test sets. To compare between baselines, we additionally report the Gini coefficient evaluated only across those languages for which we have test sets as shown in Table 8. We observe that region-specific choices (MuRIL$_{base}$ fine-tuned on HI) lead to the lowest value, similar to what we observe with the global metric.

A.4 Fine-tuning Details

We fine-tune all models using the hyperparameters mentioned in Table 9 for each task and model consistently throughout the paper. We make use of the XTREME codebase[14] to finetune these models using a NVIDIA A100 GPU. We make an exception for IndicBERT when fine-tuning on NER, where we fine-tune for 15 epochs instead of 10, to reach convergence.

A.5 Budget Allocation

In Section 4, we describe an empirical budget allocation scheme for fine-tuning of pre-trained models

that can jointly optimize on our proposed metrics. We follow a greedy approach to solve this problem, as shown in Table 10. In this paper, we solve this for one task, namely NER, but the methodology proposed is generally extensible to any task and combination of languages since it is purely empirical. We select seven source languages for which we have enough training data and fine-tune MuRIL$_{large}$ and XLM-R$_{large}$ for each of these source languages independently, for two epochs. During fine-tuning, we evaluate on each of our target languages after every 10 steps of training. Given our batch-size is 32, we gather data-points at a step size of 320 training instances. Consequently, say we have 5000 training instances for a source language, we gather approximately 30 sample points for that source language and any target language. Using these, we plot best-fit curves for $\forall (s, t)$ pairs using the *scipy.optimize.curve_fit* package. Given a function, $f(x)$, *curve_fit* uses nonlinear least squares to fit $f(x)$ to the observed data-points. We define $f(x)_{s,t} = a_{s,t} + b_{s,t} * x^{-c_{s,t}}$, because the relation between model performance and training data follows a power-law distribution (Rosenfeld et al., 2019). The best-fit curves for each source and target pair are shown in Table 11. The visualizations of the best-fit curves for a sample training language (Tamil) are shown in Figures 2, 3. Having determined constant values $\{a_{s,t}, b_{s,t}, c_{s,t}\}$ $\forall (s, t)$ independently, we proceed with finding the optimal allocation using the algorithm described in Table 10. We solve this for three different budgets, i.e., 1,000; 5,000 and 10,000 and the optimal allocations for each budget are shown in Table 12.

---

[14]https://github.com/google-research/xtreme

1775

Greedy Algorithm
1:    **Input**: Fine-tuning labeled data $\forall s \; \epsilon \; S$. A fixed budget of labeled data instances X
2:    **Initialize**: Set the total number of allocated instances to zero, i.e., $\text{allocated} = 0$, the number of allocated samples for each source language to zero, i.e. $\text{samples}[s] = 0 \forall s \; \epsilon \; S$, the current global metric for each source language to -inf, i.e. $\text{current\_gm}[s] = -\inf \forall s \; \epsilon \; S$ and the current gini coefficient for each source language to 1, i.e. $\text{current\_gini}[s] = 1 \forall s \; \epsilon \; S$
3:    **while** allocated $<$ X **do**
4:       highest_marginal_gain $= 0$
5:       **for** s in S **do**
6:          $\text{gm}_s = \sum_{t \epsilon T} d_t^{(\tau)} * (a_{s,t} + b_{s,t} * (\text{samples}[s] + 1)^{-c_{s,t}})$
7:          $\text{gini}_s = F[\text{abs}(\text{performance}_{s,t}(\text{samples}[s] + 1)) \forall t \; \epsilon \; T]$
8:          $\Delta \text{gm}_s = \text{gm}_s - \text{current\_gm}[s]$
9:          $\Delta \text{gini}_s = \text{current\_gini}[s] - \text{gini}_s$
10:         $\text{marginal\_gain} = \alpha * \Delta \text{gm}_s + \beta * \Delta \text{gini}_s$
9:         **if** marginal_gain $>$ highest_marginal_gain **do**
10:          highest_marginal_gain = marginal_gain
11:          best_language = s
12:          best_gm = gm$_s$
13:          best_gini = gini$_s$
14:         **end if**
15:       **end for**
16:       samples[s] = samples[s] + 1
17:       allocated = allocated + 1
18:       current_gm[best_language] = best_gm
19:       current_gini[best_language] = best_gini
20:    **end while**

Table 10: A greedy approach to solve the constrained optimization for the budget allocation problem as described in Appendix A.5.
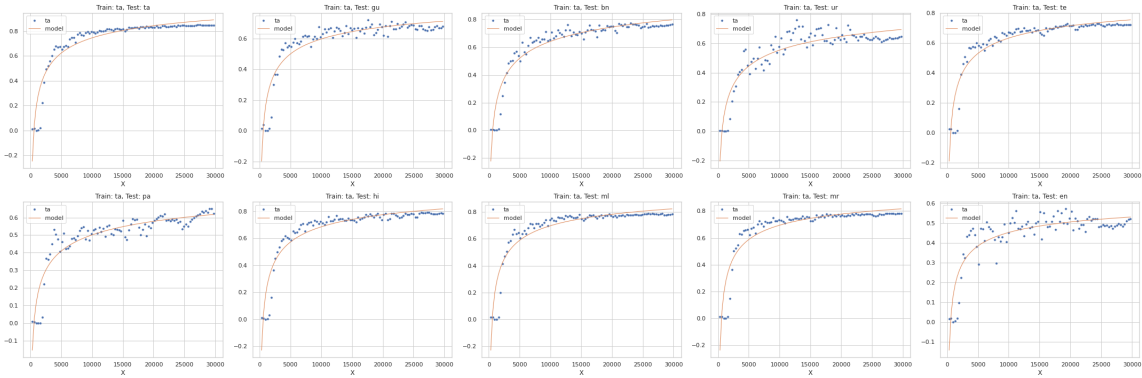

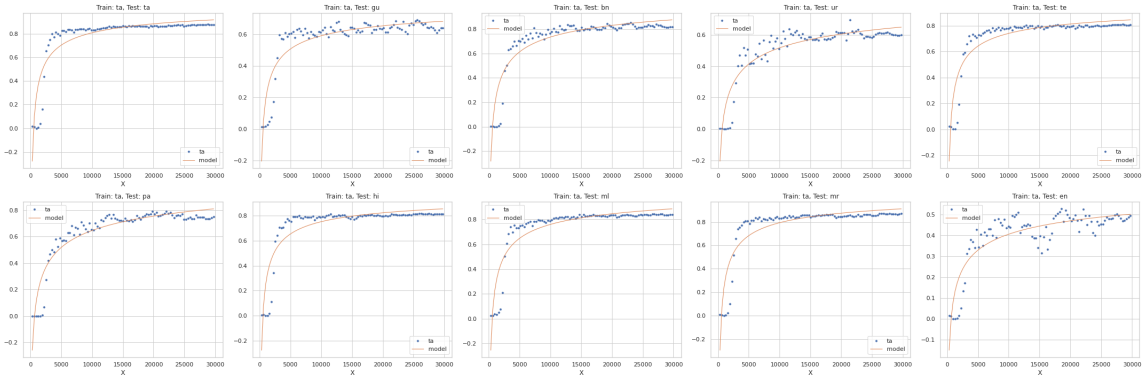Figure 2: *Best-fit curves for XLM-R* when fine-tuned on Tamil for each of the target languages.

Figure 3: *Best-fit curves for MuRIL* when fine-tuned on Tamil for each of the target languages.

Test	Train	MuRIL Edge Weight	MuRIL R-squared	XLM-R Edge Weight	XLM-R R-squared
bn	bn	$1.2 - 29.0 * x^{-0.5}$	0.88	$1.3 - 11.5 * x^{-0.4}$	0.93
	en	$1.2 - 11.4 * x^{-0.4}$	0.78	$1.1 - 8.1 * x^{-0.3}$	0.89
	hi	$1.4 - 9.4 * x^{-0.3}$	0.85	$1.1 - 8.1 * x^{-0.3}$	0.92
	ml	$1.2 - 10.7 * x^{-0.3}$	0.86	$2.3 - 4.8 * x^{-0.1}$	0.92
	mr	$1.9 - 6.5 * x^{-0.2}$	0.88	$1.9 - 4.6 * x^{-0.1}$	0.93
	ta	$1.2 - 10.5 * x^{-0.3}$	0.83	$1.3 - 6.1 * x^{-0.2}$	0.90
	ur	$1.0 - 13.5 * x^{-0.4}$	0.88	$1.0 - 6.5 * x^{-0.3}$	0.91
en	bn	$0.9 - 4.4 * x^{-0.3}$	0.86	$1.0 - 5.2 * x^{-0.3}$	0.90
	en	$1.1 - 16.4 * x^{-0.4}$	0.82	$1.1 - 14.6 * x^{-0.4}$	0.85
	hi	$1.0 - 5.6 * x^{-0.3}$	0.88	$1.0 - 7.6 * x^{-0.3}$	0.90
	ml	$1.9 - 3.5 * x^{-0.1}$	0.88	$1.0 - 6.1 * x^{-0.3}$	0.86
	mr	$1.2 - 3.2 * x^{-0.2}$	0.84	$1.2 - 4.8 * x^{-0.2}$	0.91
	ta	$0.8 - 4.2 * x^{-0.3}$	0.76	$0.7 - 6.9 * x^{-0.4}$	0.76
	ur	-	0.88	$1.0 - 3.9 * x^{-0.2}$	0.90
gu	bn	$2.6 - 4.3 * x^{-0.1}$	0.93	$1.0 - 4.8 * x^{-0.3}$	0.88
	en	$0.9 - 5.5 * x^{-0.3}$	0.80	$0.7 - 11.3 * x^{-0.5}$	0.78
	hi	$1.2 - 5.4 * x^{-0.3}$	0.87	$0.7 - 13.3 * x^{-0.5}$	0.86
	ml	$1.2 - 7.8 * x^{-0.3}$	0.85	$1.6 - 4.3 * x^{-0.2}$	0.90
	mr	$1.1 - 6.4 * x^{-0.3}$	0.87	$1.1 - 6.0 * x^{-0.3}$	0.85
	ta	$0.8 - 11.3 * x^{-0.4}$	0.78	$1.0 - 7.6 * x^{-0.3}$	0.84
	ur	$1.4 - 3.4 * x^{-0.1}$	0.91	$1.6 - 3.2 * x^{-0.1}$	0.89
hi	bn	$0.9 - 17.2 * x^{-0.5}$	0.88	$1.2 - 4.8 * x^{-0.2}$	0.94
	en	$1.0 - 11.7 * x^{-0.4}$	0.83	$0.9 - 8.6 * x^{-0.4}$	0.88
	hi	$1.1 - 23.9 * x^{-0.5}$	0.90	$1.3 - 9.5 * x^{-0.3}$	0.92
	ml	$1.1 - 12.4 * x^{-0.4}$	0.85	$1.4 - 5.7 * x^{-0.2}$	0.90
	mr	$1.1 - 17.6 * x^{-0.5}$	0.85	$2.0 - 5.3 * x^{-0.2}$	0.93
	ta	$1.0 - 19.1 * x^{-0.5}$	0.78	$1.1 - 8.7 * x^{-0.3}$	0.88
	ur	$1.0 - 8.0 * x^{-0.3}$	0.92	$1.2 - 4.6 * x^{-0.2}$	0.94
ml	bn	$1.1 - 5.9 * x^{-0.3}$	0.88	$1.3 - 4.3 * x^{-0.2}$	0.92
	en	$1.2 - 5.0 * x^{-0.2}$	0.85	$0.8 - 7.6 * x^{-0.3}$	0.85
	hi	$2.1 - 5.0 * x^{-0.1}$	0.86	$1.0 - 10.8 * x^{-0.4}$	0.90
	ml	$1.1 - 21.4 * x^{-0.5}$	0.83	$1.3 - 7.8 * x^{-0.3}$	0.90
	mr	$1.5 - 7.3 * x^{-0.3}$	0.86	$1.4 - 6.4 * x^{-0.3}$	0.91
	ta	$1.1 - 12.8 * x^{-0.4}$	0.81	$1.1 - 10.0 * x^{-0.4}$	0.86
	ur	$1.1 - 5.1 * x^{-0.2}$	0.89	$1.1 - 4.7 * x^{-0.2}$	0.89
mr	bn	$1.0 - 9.3 * x^{-0.4}$	0.88	$1.1 - 5.1 * x^{-0.2}$	0.89
	en	$0.9 - 8.8 * x^{-0.3}$	0.81	$0.9 - 7.6 * x^{-0.3}$	0.87
	hi	$1.3 - 10.3 * x^{-0.3}$	0.86	$1.1 - 9.6 * x^{-0.4}$	0.91
	ml	$1.1 - 15.2 * x^{-0.4}$	0.83	$1.3 - 6.4 * x^{-0.3}$	0.90
	mr	$1.2 - 21.9 * x^{-0.5}$	0.86	$1.6 - 7.6 * x^{-0.3}$	0.92
	ta	$1.1 - 17.3 * x^{-0.4}$	0.79	$1.1 - 10.7 * x^{-0.4}$	0.85
	ur	$1.2 - 5.2 * x^{-0.2}$	0.92	$1.3 - 4.2 * x^{-0.2}$	0.91
pa	bn	$1.0 - 6.4 * x^{-0.3}$	0.86	$0.9 - 4.2 * x^{-0.3}$	0.82
	en	$1.2 - 4.0 * x^{-0.2}$	0.84	$1.1 - 2.9 * x^{-0.2}$	0.85
	hi	$1.9 - 5.9 * x^{-0.2}$	0.84	$1.8 - 4.0 * x^{-0.1}$	0.93
	ml	$1.0 - 9.7 * x^{-0.4}$	0.83	$1.2 - 3.6 * x^{-0.2}$	0.87
	mr	$2.7 - 5.3 * x^{-0.1}$	0.88	$1.3 - 3.9 * x^{-0.2}$	0.84
	ta	$1.4 - 6.3 * x^{-0.2}$	0.86	$1.0 - 4.7 * x^{-0.2}$	0.84
	ur	$1.2 - 4.5 * x^{-0.2}$	0.92	$0.8 - 4.2 * x^{-0.3}$	0.87
ta	bn	$1.1 - 7.2 * x^{-0.3}$	0.89	$1.0 - 4.6 * x^{-0.2}$	0.93
	en	$1.0 - 7.9 * x^{-0.3}$	0.83	$0.8 - 6.7 * x^{-0.3}$	0.86
	hi	$1.4 - 6.7 * x^{-0.3}$	0.90	$1.2 - 5.9 * x^{-0.3}$	0.92
	ml	$1.0 - 14.1 * x^{-0.4}$	0.83	$1.3 - 4.7 * x^{-0.2}$	0.92
	mr	$1.3 - 9.7 * x^{-0.3}$	0.86	$2.7 - 5.0 * x^{-0.1}$	0.94
	ta	$1.1 - 19.7 * x^{-0.5}$	0.79	$1.2 - 9.4 * x^{-0.3}$	0.88
	ur	$1.2 - 5.0 * x^{-0.2}$	0.92	$1.5 - 3.4 * x^{-0.1}$	0.92
te	bn	$1.1 - 5.4 * x^{-0.3}$	0.90	$0.8 - 4.7 * x^{-0.3}$	0.88
	en	$0.8 - 10.1 * x^{-0.4}$	0.79	$0.7 - 6.7 * x^{-0.4}$	0.83
	hi	$1.0 - 9.7 * x^{-0.4}$	0.91	$0.9 - 7.3 * x^{-0.3}$	0.86
	ml	$1.0 - 15.3 * x^{-0.4}$	0.83	$1.1 - 5.6 * x^{-0.3}$	0.88
	mr	$1.0 - 12.5 * x^{-0.4}$	0.87	$1.7 - 4.5 * x^{-0.2}$	0.93
	ta	$1.0 - 16.5 * x^{-0.4}$	0.81	$1.0 - 7.7 * x^{-0.3}$	0.87
	ur	$1.4 - 4.2 * x^{-0.2}$	0.91	$1.4 - 3.1 * x^{-0.1}$	0.90
ur	bn	$0.6 - 11.3 * x^{-0.5}$	0.86	-	0.76
	en	$1.1 - 5.5 * x^{-0.2}$	0.83	$1.0 - 5.9 * x^{-0.3}$	0.81
	hi	$2.6 - 4.8 * x^{-0.1}$	0.85	-	0.96
	ml	$1.1 - 8.2 * x^{-0.3}$	0.80	$2.5 - 5.0 * x^{-0.1}$	0.85
	mr	$4.3 - 6.1 * x^{-0.1}$	0.87	$5.2 - 7.0 * x^{-0.0}$	0.91
	ta	$1.1 - 5.0 * x^{-0.2}$	0.83	$1.2 - 5.2 * x^{-0.2}$	0.83
	ur	$1.0 - 42.2 * x^{-0.6}$	0.87	$1.1 - 20.9 * x^{-0.5}$	0.90

Table 11: *Power-law equations* empirically determined for each source and target pair. Please refer to Section A.5 for more details

Metric	Budget	Model	bn	en	hi	ml	mr	ta	ur
$GM_{\tau=0}$	1,000	XLM-R$_{large}$	128	157	145	134	133	163	140
		MuRIL$_{large}$	137	135	134	158	142	159	135
	5,000	XLM-R$_{large}$	704	792	693	794	696	628	693
		MuRIL$_{large}$	743	644	749	783	745	852	484
	10,000	XLM-R$_{large}$	1322	1349	1400	1481	1457	1479	1512
		MuRIL$_{large}$	1302	1468	1379	1421	1425	1448	1557
$GM_{\tau=1}$	1,000	XLM-R$_{large}$	126	160	159	134	129	163	129
		MuRIL$_{large}$	142	136	152	143	148	157	122
	5,000	XLM-R$_{large}$	710	805	713	803	707	639	623
		MuRIL$_{large}$	744	644	761	772	747	848	484
	10,000	XLM-R$_{large}$	1308	1363	1456	1465	1459	1471	1478
		MuRIL$_{large}$	1308	1488	1396	1406	1416	1441	1545

Table 12: *Optimal allocations under different budgets.* Please refer to Section A.5 for more details

Metric	Train Lang.	Model	NER	POS	NLI	QA	Average
$M_{\tau=0}$ ↑ (Linguistic)	English	IndicBERT	16.5	16.1	6.5	5.3	11.1
		XLM-R$_{base}$	27.0	21.4	10.3	13.8	18.1
		MuRIL$_{base}$	33.4	20.7	10.5	14.9	19.9
		XLM-R$_{large}$	28.7	21.9	11.0	15.6	19.3
		MuRIL$_{large}$	31.5	21.3	11.1	**15.9**	20.0
	Hindi	IndicBERT	23.7	17.6	6.6	4.8	13.2
		XLM-R$_{base}$	30.4	22.4	10.6	13.5	19.2
		MuRIL$_{base}$	**34.0**	**22.7**	10.8	14.7	20.6
		XLM-R$_{large}$	33.0	22.4	**11.5**	15.2	20.5
		MuRIL$_{large}$	33.4	22.4	11.4	15.7	**20.7**
$M_{\tau=1}$ ↑ (Demographic)	English	IndicBERT	39.2	44.2	36.6	28.4	37.1
		XLM-R$_{base}$	59.2	58.1	43.6	49.9	52.7
		MuRIL$_{base}$	69.6	54.7	45.5	53.8	55.9
		XLM-R$_{large}$	61.2	60.3	46.6	56.6	56.2
		MuRIL$_{large}$	68.2	58.6	47.4	**57.9**	58.0
	Hindi	IndicBERT	61.0	61.6	39.8	29.9	48.1
		XLM-R$_{base}$	70.3	66.7	45.8	50.6	58.3
		MuRIL$_{base}$	**75.1**	**67.3**	46.8	54.7	61.0
		XLM-R$_{large}$	74.4	66.8	**49.4**	53.2	60.9
		MuRIL$_{large}$	74.8	66.5	49.2	54.6	**61.3**
Gini Coeff. ↓	English	IndicBERT	0.67	0.81	0.92	0.84	0.81
		XLM-R$_{base}$	0.61	0.76	0.88	0.83	0.77
		MuRIL$_{base}$	0.59	0.76	0.88	0.83	**0.76**
		XLM-R$_{large}$	0.6	0.75	0.88	0.83	0.77
		MuRIL$_{large}$	0.59	0.76	0.88	0.83	0.77
	Hindi	IndicBERT	0.68	0.8	0.91	0.83	0.81
		XLM-R$_{base}$	0.59	0.75	0.87	0.83	**0.76**
		MuRIL$_{base}$	0.59	0.75	0.87	0.83	**0.76**
		XLM-R$_{large}$	0.59	0.76	0.88	0.83	0.77
		MuRIL$_{large}$	0.59	0.75	0.87	0.83	**0.76**
Efficiency ↑	English	IndicBERT	53.6	50.2	56.7	66.0	56.6
		XLM-R$_{base}$	44.4	48.1	57.4	76.7	56.7
		MuRIL$_{base}$	69.1	52.5	58.7	**77.8**	64.5
		XLM-R$_{large}$	44.4	48	57.3	76.3	56.5
		MuRIL$_{large}$	63.1	51.8	57.8	75.7	62.1
	Hindi	IndicBERT	62.5	53.7	56.9	64.5	59.4
		XLM-R$_{base}$	48.3	50.0	58.5	75.9	58.2
		MuRIL$_{base}$	**69.8**	**56.2**	**59.8**	77.3	**65.8**
		XLM-R$_{large}$	49.2	49.0	59.1	75.1	58.1
		MuRIL$_{large}$	65.2	53.7	58.8	75.0	63.2

Table 13: *Zero-shot fine-tuning results.* Overall, MuRIL$_{large}$ scores highest on the utility metrics, the Gini coefficient is relatively high across all models and both MuRIL$_{base}$ and MuRIL$_{large}$ are, on average, equal with regards to efficiency. Note that the metrics are computed considering all 23 languages as detailed in Section 5.1. More discussions in Section 5.2.