Uplifting Lower-Income Data: Strategies for Socioeconomic Perspective Shifts in Large Multi-modal Models

Joan Nwatu[1] and **Oana Ignat**[2] and **Rada Mihalcea**[1]
[1]University of Michigan, USA
[2]Santa Clara University, USA
{jnwatu, mihalcea} @umich.edu   oignat@scu.edu

Abstract

Recent work has demonstrated that the unequal representation of cultures and socioeconomic groups in training data leads to biased Large Multi-modal (LMM) models. To improve LMM model performance on underrepresented data, we propose and evaluate several prompting strategies using non-English, geographic, and socioeconomic attributes. We show that these geographic and socioeconomic integrated prompts favor retrieving topic appearances commonly found in data from low-income households across different countries leading to improved LMM model performance on lower-income data. Our analyses identify and highlight contexts where these strategies yield the most improvements.

1 Introduction

A lack of diversity in popular AI datasets (Shankar et al., 2017) leads to unequal model performance, further widening the technological gap between well-represented and underrepresented communities. While data from higher-income Western communities are readily available online, lower-income and non-Western data are often missing (Rosling et al., 2019). As a result, cost-effective methods like web scraping fail to produce diverse datasets.

One approach to building large datasets leverages LMM models to filter uncurated data based on image-text association strength scores (Fang et al., 2023). For instance, OpenAI's CLIP ViT-B/32 (Radford et al., 2021) was used to filter web-scraped images to create the LAION-5B dataset (Schuhmann et al., 2022). However, foundation LMM models like CLIP perform unequally across cultures and socioeconomic groups, favoring higher-income and Western images (Nwatu et al., 2023).

Datasets filtered by LMM models reflect the model's biases (Fang et al., 2023), often excluding underrepresented data and worsening the lack of diversity in AI models. Ignat et al. (2024) demonstrates this by showing that the LAION-5B dataset closely resembles data from Western countries, such as the United States and Canada while differing from non-Western countries' data. This leads to LMM models with uneven performance on data drawn from different locations and income groups. Therefore, our paper seeks to answer the following question: ***How do we improve the performance of LMM models on lower-income and non-Western data?***

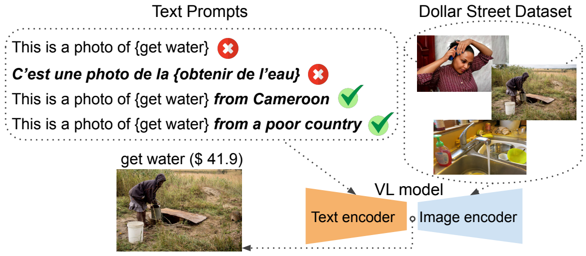

Figure 1: Low-income Image Retrieval from Dollar Street dataset (Rojas et al., 2022) using different prompt formulations. Prompts with integrated *country* and *income* information successfully retrieve fewer standard images previously left out by the English and translated (French) prompts.

We tackle performance inequality in LMM models (Radford et al., 2021; Visheratin, 2023) through prompting that transfers the cultural knowledge embedded in language (Ventura et al., 2023; Buettner et al., 2024; Nguyen et al., 2024). Our goal is to improve the performance of LMM models on data from households with non-Western and lower socioeconomic status. Specifically, as shown in Fig. 1, we pose several research questions to evaluate the role of non-English languages, as well as prompts with geographic and socioeconomic attributes, to retrieve more diverse images.

Our contributions are summarized as follows. First, we show that a naive **prompt translation-based approach fails** to adequately address the performance gap of LMM models on lower-income

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2127–2144
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

data. Second, we establish that **geographic and socioeconomic attribute integrated prompts improve** LMM performance on lower-income data. We identify contexts where these prompts work best by conducting an in-depth analysis of LMM models' understanding of these attributes and their effects on recall across data from different countries. Lastly, we share insights from our analysis demonstrating how **these attributes drive a perspective shift** that benefits the retrieval of lower-income data.

2 Related Work

Addressing AI Performance Inequality. Class imbalances in training data contribute significantly to bias in AI models (Ferrara, 2024; Shankar et al., 2017; He and Garcia, 2009; Pouget et al., 2024), leading to unequal outcomes in areas like facial recognition (Buolamwini and Gebru, 2018), healthcare (Obermeyer et al., 2019), and hiring (Raghavan et al., 2020). Since creating balanced datasets is challenging and costly (Ignat et al., 2024; Ramaswamy et al., 2023), researchers have explored bias mitigation techniques such as data augmentation, feature importance tuning, regularization, and adversarial training (Yan et al., 2020; Zafar et al., 2017; Ignat et al., 2024; Maudslay et al., 2019; Sharma et al., 2020; Navarro et al., 2024; Zhang et al., 2018). Our work is most similar to research on post-processing methods Ferrara (2023); Hardt et al. (2016); Kamiran et al. (2012); Pleiss et al. (2017) that adjust model outcomes to meet diversity standards, aiming to benefit disadvantaged groups. Prior research has shown that LMM models perform poorly on data from lower socioeconomic groups, and our analysis investigates non-invasive post-processing methods to address this issue.

Multilingual AI Models. Language plays a key role in transmitting cultural knowledge (Callies, 2024; Sharifian, 2014; Karsdorp and Fonteyn, 2019; Norton, 1997), as AI models often absorb biases from the language in their training data (Stanczak and Augenstein, 2021; Rogers et al., 2021) and model outputs can be controlled by specifying a cultural shift in perspective (Ventura et al., 2023) to improve diversity. However, research Arora et al. (2023); Cao et al. (2023); AlKhamissi et al. (2024); Liu et al. (2021) shows that large language models (LLMs) and LMM models capture more cultural information from English data (mainly Western) than from non-English data. This disparity stems from differences in the quantity and quality of non-English data, translation issues, and model design (Arora et al., 2023; Hershcovich et al., 2022; Nasif et al., 1991).

Similar to past studies (De Vries et al., 2019; Nguyen et al., 2024) using multilingual approaches to enhance data diversity, our work explores how multilingual large multi-modal models and non-English languages can improve representation across regions and income groups.

Prompting AI Models. Recent studies have explored prompting techniques for large language models, including both hard (Petroni et al., 2019; Zhou et al., 2023) and soft prompting (Huang et al., 2023; Goswami et al., 2023), to improve model adaptation for tasks like instruction tuning, and value alignment. These methods are also applied in LMM models (Lu et al., 2022; Yao et al., 2024; Zhou et al., 2022). While prior work (Buettner et al., 2024) has incorporated geographic and physical attributes into prompts to enhance image retrieval diversity, this research extends the investigation to non-English language prompts and socioeconomic attributes to analyze how LMM models encode representations of various topics across regions and socioeconomic status.

3 Methodology

We propose prompting strategies that account for language, location, and socio-economic attributes and analyze how these prompts affect the performance of a multilingual LMM model on data across different socio-economic groups, primarily focusing on lower-income data.

3.1 Dollar Street Dataset

We use the Dollar Street (Rojas et al., 2022), which contains $38,479$ images of household items (e.g., "stoves", "cutlery", "toothbrush") spanning a large number of countries and several income levels. The dataset images were sourced from households in 63 countries on four continents (Africa, America, Asia, and Europe). The number of images ranges from 45 in Canada to $4,704$ in India, with a median of 407 images per country. Size and image resolutions vary slightly across data from different regions; however, the mean and median image properties per region are relatively similar.

Image Income Classes. Each image is accompanied by the monthly household income value in U.S.

dollars, calculated to reflect monthly consumption and adjusted for purchasing power parity to match the variance in cost of living across the different regions. The monthly income values range from $26.9\$$ to $19,671.0\$$.

For fair comparison across bins, we group the images using the quartile binning method, which splits the data into an approximately equal number of images per bin as shown in Rojas et al. (2022). We group the images into four income classes (“poor”, “low-mid”, “up-mid”, and “rich”) using quartiles as shown in Table 1. We further categorize the lowest two image income classes as *lower-income images* and the highest two income groups as *higher-income images*.

Quartile name	Income range
poor	26.9 - 95.0
low-mid	195.4 - 685.0
up-mid	694.0 - 1,998.0
rich	2,001.0 - 19,671.0

Table 1: Income quartiles and their ranges for all the images in Dollar Street.

Country Economic Classes. We group all 63 countries from Dollar Street into country economic classes based on their World Bank income classification.[1] All the countries and their economic classes are shown in Appendix A.1. We further categorize the lowest two country economic classes as *lower-income countries* and the highest two economic groups as *higher-income countries*.

Topic Representations. There are 291 unique topics associated with the images in the dataset which reflect everyday household objects and human actions (e.g., “toilet paper”, “get water”), some of which are subjective (e.g., “next big thing I plan to buy”, “favorite sports clubs”, “most loved item”). We remove nineteen subjective topics from the dataset following De Vries et al. (2019) and Nwatu et al. (2023).

3.2 Prompt Design

We describe below the prompting strategies we use for our experiments and show examples in Figure 1.

Default English Topic Prompt. Using the topics, we formulate an English prompt without any modifications (e.g., “This is a photo of *cutlery*”), as described in Radford et al. (2021), to which we refer to as the *default English prompt*. The performance obtained using these prompts is set as our baseline.

Translated Topic Prompt. For our multilingual experiments, we investigate the impact of non-English language prompts on the Dollar Street dataset. We use the term *non-English major language* to refer to the non-English language that is most widely spoken or most commonly used in a particular country or region.

Specifically, we pair each country with their non-English major language (e.g., *Portuguese* for *Brazil*, *French* for *Cameroon*) following the country and language information provided by official sources.[2]

We identify 59/63 countries in Dollar Street where one or more major non-English languages are spoken. We also select languages covered by state-of-the-art machine translation and multilingual LMM models. There are 40 such non-English major languages, and they are listed in Appendix A.1.

Finally, we translate the *default English prompts* to these 40 languages using the NLLB-200-distilled-600M (Costa-jussà et al., 2022), an open-source state-of-the-art neural machine translation model. Translation metrics for NLLB-200-distilled-600M are shown in Appendix Table 13 and available on HuggingFace. If an image prompt is translated into the non-English major language of the image’s country of origin, it is referred to as a *native translated prompt*.

Country Suffix Topic Prompt. For our second prompting technique, we include country names as suffixes to the default English prompt (e.g., “This is a photo of *cutlery* from *Cameroon*”). We create 63 new prompt templates by adding the country names of each of the 63 countries in Dollar Street. We refer to these prompts as *country-suffix prompts*.

Income Suffix Topic Prompt. We also create prompts by integrating socio-economic attributes (e.g., “poor country”, “rich region”) as suffixes to the *default English prompt*. For instance, a sample prompt is “This is a photo of *cutlery* from *a rich country*”. For more robust results, we use multiple synonyms each for the *poor* and *rich* attributes (e.g., “an impoverished country”, “a wealthy region”).

[1] https://datahelpdesk.worldbank.org/

[2] www.cia.gov/the-world-factbook/field/languages/, www.ncsc.org/__data/assets/pdf_file/0024/17862/languagesbycountries.pdf, www.dss.gov.au/sites/default/files/files/foi_disclosure_log/12-12-13/language-list.pdf

We also create prompts using neutral suffixes (e.g., "a country", "a home"). We refer to these prompts as *income-suffix prompts*.

3.3 State-of-the-art LMM Model

For our evaluation, we chose NLLB-CLIP-SigLIP (Visheratin, 2023), a state-of-the-art multilingual LMM model, with broad reach across many low-resource languages and superior performance among other models.[3] The model consists of an image encoder from the SigLIP model (Beyer et al., 2022; Zhai et al., 2023) and a text encoder from the NLLB model (Costa-jussà et al., 2022). The model supports the 201 languages of the Flores-200 (Costa-jussà et al., 2022) and has recorded ground-breaking results on the Crossmodal-3600 dataset (Thapliyal et al., 2022), especially on low-resource languages.

4 Research Questions

We perform several analyses to answer three research questions that uncover and mitigate limitations in the performance of LMM models across different countries and socioeconomic groups.

4.1 RQ1. Do translated prompts improve retrieval performance for lower-income images?

We calculate the cosine similarities between image and translated prompt text embeddings for each image-topic pair across English and 40 non-English languages, generating 41 alignment scores per image. The alignment scores with *default English prompts* serve as our baseline.

We compute Recall scores by selecting the top N images with the highest alignment scores for each topic, where N represents the number of ground truth images. We then group and analyze the Recall scores across different countries and image income classes and present our findings below.

***Native translated prompts* perform consistently worse than *English prompts* on lower-income images from their respective countries.** We focus our analysis on images from the two lowest image income groups, i.e., *poor* and *low-middle* as grouped in 3.1. After excluding 20 countries without data for these income groups (e.g., *Russia*, *Turkey*), we retain 39/59 countries and 28/40 non-English languages for the study. Each country is paired with its native non-English language, and we compare Recall scores for the *native translated prompts* to those for the *default English prompts*. The average Recall across all countries and scores from four countries are displayed in Figure 2.

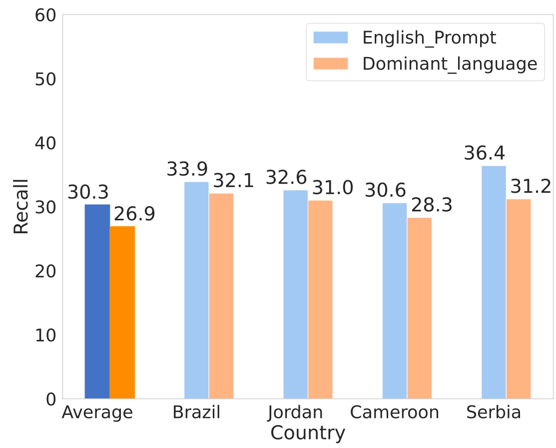

Figure 2: NLLB SigLIP Recall (%) over poor and lower-middle income images from four countries, one from each of the four continents: Asia, Africa, America, and Europe for English and native translated prompts. *Best viewed in color.*

For 35 out of 39 countries, the *native translated prompts* underperform compared to the *default English prompts*. The exceptions include *Burkina Faso*, *Nigeria*, *Pakistan* and *Tanzania*, where *native translated prompts* in *French* (diff. of 1.0), *Hausa* (diff. of 0.2), *Urdu* (diff. of 0.7) and *Swahili* (diff. of 1.5), respectively, outperform English prompts. Overall, native translated prompts generally fail to retrieve diverse images, as depicted by the example using French prompt in Fig. 1.

The best-performing non-English language often differs from the country's native language. We analyze the Recall scores for lower-income images across 28 language prompts used in different countries and find that the best-performing language prompts often differ from the countries' non-English major languages. Specifically, in 24 out of 39 countries, non-English language prompts outperform the default English prompts, yet these top-performing languages are not typically spoken in the respective countries. As illustrated in Figure 3, for 37/39 countries, the language with the highest Recall score (highlighted in yellow) differs from the country's primary non-English language (highlighted in cyan), with exceptions in Indonesia and Pakistan, where they coin-

[3] https://huggingface.co/visheratin/nllb-clip-large-siglip

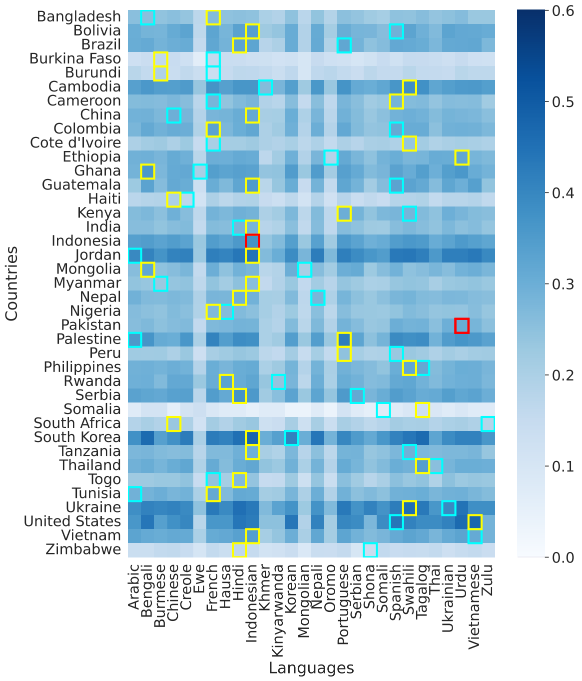

Figure 3: Recall scores for *lower* income images from 39 countries and 28 languages. The cyan highlight shows the Recall for a country's native translated language, the yellow highlight shows the best-performing language recall, and the red shows the Recall for the language that is both the native and highest performing for that country. *Best viewed in color.*

cide (highlighted in bold red).

***Translated prompts* decrease performance for all image income classes across all countries.** We analyze the impact of 40 non-English language prompts on all images from Dollar Street, covering 59 countries and we group Recall scores by image income classes. By comparing Recall scores between *default English prompts* and *native translated prompts*, we assess the effect of each non-English language on four income classes and show the difference in scores in Appendix Table 10.

Non-English languages (Average)	Image Income Class			
	Poor Δ	**Low-mid Δ**	**Up-mid Δ**	**Rich Δ**
	20.2 (-2.2)	31.0 (-4.9)	37.8 (-7.8)	36.1 (-7.5)

Table 2: Average differences between Recall scores for non-English language prompts and Recall scores for default English prompts for all data, grouped by image income classes. We find that non-English prompts lead to a decrease in Recall scores across all income classes.

We show in Table 2 the average Recall and drops in performance across all 40 translated prompts for each image income class. The results indicate that higher-income classes, specifically the *rich* and *up-mid* groups, experience the largest drops in performance with translated prompts. This may be due to the overrepresentation of images from these income groups in AI models and datasets, positioning them as the "standard" representation. Similarly, English, the dominant training language, is seen as the "standard" for textual data, so non-English prompts may signal a deviation from this standard, resulting in poorer model performance.

4.2 RQ2. Does adding country information improve retrieval performance for lower-income images?

We compute cosine similarity scores between NLLB-CLIP-SigLIP image embeddings and the text embeddings of 63 *country suffix prompts.*, yielding 63 image-topic alignment scores per image. Using the alignment scores from the *default English prompts* as a baseline, we follow the procedure outlined in Section 4.1 to calculate Recall scores for each topic with the *country suffix prompts*. We then analyze the impact of adding country suffixes to text prompts and present the results in the following sections.

***Country-suffix prompts* perform consistently better than *default English prompts* on lower-income images.** Focusing on low-income data, we filter out 21 countries without images from *poor* or *low-mid* income households, leaving 42 countries for analysis. In Figure 4, we present the average Recall scores across all countries using both *default English* and *country-suffix prompts*, along with results from four sample countries from different continents.

Our findings indicate that in 38/42 countries, adding a *country-suffix* to text prompts improves Recall performance for lower-income images compared to *default English prompts*. Exceptions include *Bolivia, Brazil, Jordan*, and the *United States*. *Country-suffix prompts* are thus more effective in retrieving diverse images, as demonstrated by the Cameroon example in Fig. 1.

A country's economic status influences the performance of its *country-suffix prompt* across different image income classes. Country suffix prompts improve LMM model performance for lower-income images (*poor*) but reduce performance for higher-income images. Using World Bank income classifications, we calculate Recall

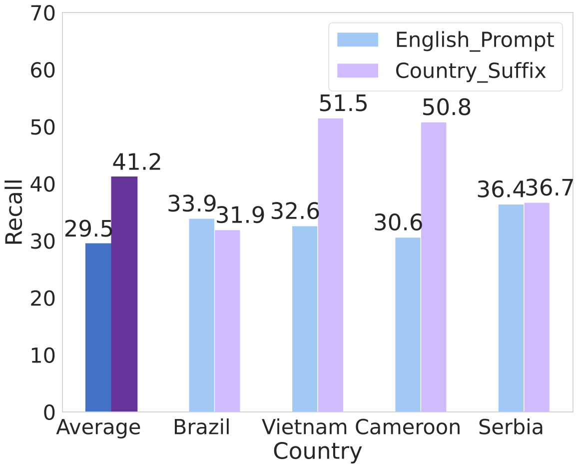

Figure 4: Recall (%) with NLLB SigLIP over *poor* and *lower-middle* income images from four countries from Asia, Africa, America, and Europe, for English and Country Suffix prompts. *Best viewed in color.*

scores across four country suffix groups (poor, low-mid, up-mid, and rich) and four image income classes (based on household income). For each image income class, we aggregate Recall scores and compare them with those from default English prompts, as shown in Table 3, with detailed results in Appendix Table 11 and Table 12. For example, Table 3 shows that Recall of images from poor households using *country suffixes* of poor countries is 31.2, a 9.7 increase from *default English prompt* performance on that group.

The analysis reveals that country suffixes from poor, low-mid, and up-mid income categories improve Recall for images from poor households, while reducing Recall for higher-income groups (low-mid, up-mid, rich).

Country suffix (Avg)	**Image Income Classes**			
	Poor Δ	**Low-mid Δ**	**Up-mid Δ**	**Rich Δ**
Poor	**31.2** (+9.7)	30.7 (-5.3)	25.5 (-20.6)	21.9 (-22.5)
Low-mid	29.2 (+4.2)	**31.6** (-4.3)	27.6 (-15.3)	23.2 (-16.9)
Up-mid	24.1 (+2.4)	31.2 (-3.7)	**32.8** (-12.9)	30.3 (-14.7)
Rich	20.8 (-2.1)	0.329 (-4.4)	**41.4** (-6.6)	40.0 (-5.6)

Table 3: Average NLLB SigLIP Recall scores for each category and the average difference between default English prompt and country suffix Recall across the four different income groups, grouping country suffixes into income class categories based on their World Bank economic classification. Recall increase is shown in green while Recall drops are highlighted in red. *Best viewed in color.*

Interestingly, country suffixes tend to favor image retrieval from income groups that match or are close to their own economic classification. When the four income classes are re-categorized into two classes (lower-income: poor, low-mid and higher-income: up-mid, rich), we find that in 48/63 cases, the image income category with the highest Recall corresponds to the country suffix's economic class, demonstrating the alignment between income levels of country-suffixes and retrieval performance.

The best-performing country suffixes for lower-income images from a continent are from the same continent. We calculate Recall results for lower-income images from 42 countries using the 42 country suffix prompts, yielding a total of 1,764 Recall scores. Using country-suffixes, we group these scores by continent and further categorize them based on the World Bank Income classes of the respective country-suffixes. We present the average Recall and differences compared to default English prompts for each group in Table 4.

Image by Continent	**Country by Income**	**Country Suffix**			
		Africa Δ	**America Δ**	**Asia Δ**	**Europe Δ**
Africa	Poor	36.6 (+15.7)	27.2 (+6.3)	23.0 (+2.1)	19.7 (-1.2)
	Low-mid	37.3 (+10.3)	31.2 (+4.2)	25.7 (-1.4)	24.3 (-2.7)
	Up-mid	24.3 (+2.1)	21.6 (-0.6)	17.2 (-5.0)	18.1 (-4.1)
	Average	**32.7 (+9.4)**	**26.7 (+3.3)**	**22.0** (-1.4)	**20.7** (-2.7)
America	Low-mid	24.5 (-2.6)	26.8 (-0.4)	20.7 (-6.5)	22.4 (-4.8)
	Up-mid	23.4 (-11.0)	35.2 (+0.9)	23.4 (-10.9)	28.9 (-5.4)
	Rich	20.4 (-15.5)	30.4 (-5.5)	22.8 (-13.1)	26.3 (-9.6)
	Average	**22.8** (-9.7)	**30.8 (-1.7)**	**22.3** (-10.2)	**25.9** (-6.6)
Asia	Low-mid	29.4 (-2.4)	30.7 (-1.1)	32.7 (+0.8)	27.9 (-3.9)
	Up-mid	28.1 (-5.1)	32.5 (-0.6)	34.6 (+1.5)	30.3 (-2.8)
	Rich	31.0 (-14.0)	33.3 (-11.7)	36.0 (-9.0)	39.6 (-5.4)
	Average	**29.5** (-7.2)	**32.2** (-4.5)	**34.4 (-2.2)**	**32.6** (-4.0)
Europe	Low-mid	19.6 (-23.7)	29.3 (-14.0)	26.4 (-16.9)	44.3 (+1.0)
	Up-mid	20.4 (-16.0)	26.7 (-9.7)	23.1 (-13.3)	35.9 (-0.5)
	Average	**20.0** (-19.9)	**28.0** (-11.9)	**24.8** (-15.1)	**40.1 (+0.3)**

Table 4: Average NLLB SigLIP Recall scores for each continent and the average difference between default English prompt Recall and country suffix Recall from the four continents, grouping lower-income images according to continents and further into groups of countries arranged by income class categories based on their World Bank economic classification. *Best viewed in color.*

Our findings emphasize the significance of regional specificity in data collection, as the best-performing suffixes align with their respective continents (shown by the diagonal of bold values in Table 4). The results indicate that lower-income images from African nations benefit significantly from including country suffixes. In contrast, data from America and Asia show no Recall improvements, underscoring the necessity for tailored data collection strategies across regions. Notably, lower-income images from African countries exhibit a Recall score of 36.6, reflecting the highest performance increase of 15.7 when using African suffixes.

The positive impact of country suffix prompt additions is particularly pronounced for Africa, as the prompts enhance performance on underrepresented data by shifting model inference from its learned standard. This effect is crucial given the current datasets often lack representation from African countries and poor households. Additionally, the similarities among African countries contribute to this improved performance.

Meanwhile, we find no Recall enhancements for higher-income data, regardless of the alignment between images and country suffixes (see Table 5).

Image by Continent	Country by Income	Country Suffix			
		Africa Δ	**America Δ**	**Asia Δ**	**Europe Δ**
Africa	Poor	42.3 (-1.4)	40.3 (-3.4)	36.0 (-7.7)	40.3 (-3.4)
	Low-mid	41.8 (-8.8)	39.7 (-10.9)	37.0 (-13.6)	40.7 (-9.9)
	Up-mid	33.3 (-5.9)	32.4 (-6.8)	25.7 (-13.5)	33.4 (-5.8)
	Average	**39.1 (-5.4)**	**37.5** (-7.0)	**32.9** (-11.6)	**38.1** (-6.4)
America	Up-mid	24.3 (-20.8)	37.6 (-7.5)	26.9 (-18.3)	37.2 (-8.0)
	Rich	28.7 (-33.4)	45.7 (-16.4)	32.5 (-29.6)	44.7 (-17.4)
	Average	**26.5** (-27.1)	**41.7 (-12.0)**	**29.7** (-24.0)	**41.0** (-12.7)
Asia	Low-mid	33.8 (-17.2)	38.4 (-12.6)	40.4 (-10.5)	43.4 (-7.6)
	Up-mid	30.9 (-21.0)	40.5 (-11.4)	41.2 (-10.7)	46.5 (-5.4)
	Rich	28.2 (-19.9)	36.1 (-12.0)	36.5 (-11.6)	40.1 (-8.0)
	Average	**31.0** (-19.4)	**38.3** (-12.0)	**39.4** (-10.9)	**43.3 (-7.0)**
Europe	Low-mid	22.3 (-23.9)	33.3 (-13.1)	26.6 (-19.6)	42.2 (-0.4)
	Up-mid	18.9 (-25.0)	29.2 (-14.7)	24.5 (-19.4)	40.7 (-3.2)
	Rich	19.0 (-20.8)	27.8 (-12.0)	21.5 (-18.3)	37.6 (-2.2)
	Average	**20.1** (-23.2)	**30.0** (-13.3)	**24.2** (-19.1)	**40.2 (-3.1)**

Table 5: Grouping higher income images according to continents and further into groups of countries arranged by income class categories based on their World Bank economic classification, this table shows the average NLLB SigLIP Recall scores for each continent and the average difference between default English prompt Recall and country suffix Recall from the four continents.

4.3 RQ3. Does adding income information improve retrieval performance for lower-income images?

We create three categories of income suffixes, *poor*, *rich*, and *neutral*, as described in Section 3.2. We repeat the image retrieval experiments from previous research questions to determine the Recall for images from each topic. We group and analyze these results across countries and income groups.

***Poor* income suffixes yield the best performance on most lower-income images.** Our analysis reveals that the *poor* income suffix prompt achieves the highest performance in 26/42 countries with lower-income images. In 12/42, default English prompts outperform all income suffixes. Nevertheless, most (30/42) countries show Recall improvements when using one of the income suffixes.

We illustrate in Figure 5 the aggregate the average Recall scores for all 42 countries across default English and the income suffix prompts. Notably, the *poor* income suffix demonstrates the best Recall, effectively retrieving a diverse array of images, as shown by the example in Fig. 1. Recall scores for four sample countries are in Appendix Figure 7.

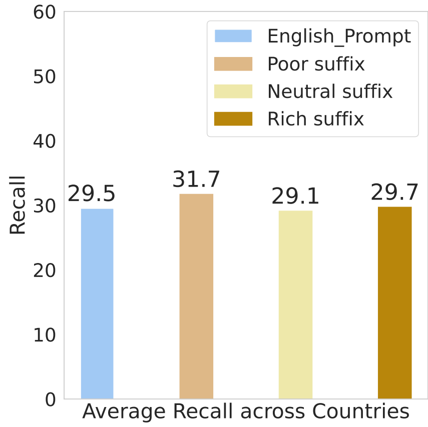

Figure 5: Average Recall with NLLB SigLIP over *poor* and *lower-middle* income images, for English and Income Suffix prompts. *Best viewed in color.*

Images from the *poor* income group benefit the most from income suffixes. We group the data into four income groups (by household income) and further categorize them according to the World Bank income classification of their country of origin. In Table 6, we show the Recall scores and performance improvements relative to the default English prompts for each data group.

We find that income suffixes predominantly benefit data from poor households and some from low-mid income households, while data from other income groups do not show Recall increases.

An interesting finding is that all income suffixes, including *rich* and *neutral*, result in decreased Recall for higher-income images (i.e., *up-mid* and *rich*). This suggests that default English prompts yield the best results for higher-income images, likely due to their high representation in AI models and datasets as the "standard representation." Consequently, the inclusion of socioeconomic status information may lead the model to prioritize lower-income images over higher-income ones. This phenomenon is evident in the results, which show Recall improvements for lower-income images while diminishing Recall for higher-income images, potentially indicating a shift in the model's perspective away from its default understanding of the topic.

Images by Income	Country by Income	Income Suffix		
		Poor ∆	**Rich ∆**	**Neutral ∆**
Poor	Poor	26.8 (+4.4)	21.9 (-0.5)	20.7 (-1.7)
	Low-mid	30.0 (+7.6)	26.0 (+3.6)	24.6 (+2.2)
	Up-mid	31.9 (+9.5)	28.0 (+5.6)	26.3 (+3.9)
	Average	**29.6** (+7.2)	**25.3** (+2.9)	**23.9** (+1.5)
Low-mid	Poor	33.3 (-2.6)	33.3 (-2.6)	33.5 (-2.4)
	Low-mid	36.5 (+0.6)	35.6 (-0.3)	35.7 (-0.2)
	Up-mid	30.6 (-5.3)	30.5 (-5.4)	30.3 (-5.6)
	Rich	35.5 (-0.5)	38.1 (+2.2)	37.2 (+1.3)
	Average	**34.0** (-2.0)	**34.4** (-1.5)	**37.2** (-1.7)
Up-mid	Poor	31.4 (-14.2)	36.4 (-9.2)	32.8 (-12.8)
	Low-mid	37.6 (-8.1)	42.4 (-3.2)	44.5 (-1.1)
	Up-mid	33.0 (-12.6)	38.0 (-7.6)	41.3 (-4.3)
	Rich	28.4 (-17.4)	33.4 (-12.2)	36.3 (-9.3)
	Average	**32.6** (-13.1)	**37.6** (-8.1)	**38.7** (-6.9)
Rich	Poor	29.6 (-14.0)	42.6 (-1.0)	42.4 (-1.2)
	Low-mid	30.6 (-13.0)	38.0 (-5.6)	38.5 (-5.1)
	Up-mid	25.8 (-17.8)	33.8 (-9.8)	36.1 (-7.5)
	Rich	25.5 (-18.1)	33.8 (-9.8)	36.5 (-7.1)
	Average	**27.9** (-15.7)	**37.1** (-6.6)	**38.4** (-5.2)

Table 6: Average NLLB SigLIP Recall scores for each category and the average difference between default English prompt Recall and income suffix recall, grouping images according to household income level and separating countries into income class categories based on their World Bank economic classification.

Model	Suffix Prompts				
	English	Country	Poor	Rich	Neutral
NLLB SigLIP	30.3	41.6	32.1	30.2	29.5
Sentence Transformers MCLIP	22.2	44.9	28.6	26.7	25.2
Open AI CLIP ViT 32/B	25.4	45.0	30.0	28.1	28.3

Table 7: Average Recall over lower-income images across 39 countries for English, Country Suffix, and three Income Suffix prompts for three LMM models

4.4 Results Significance and Generalizability

We conducted the Wilcoxon Signed Rank (Woolson, 2005) test (p-value < 0.05) to assess the statistical significance of our findings. The results indicated that the differences between the default English prompt results and each prompt intervention were statistically significant, except for the 'rich' and 'neutral' income suffix prompts (more details in Appendix Table 14).

Although our primary focus is on the NLLB-CLIP-SigLIP results, we confirmed that these findings are consistent across the two other LMM models we tested (Open AI's CLIP ViT B/32 (Radford et al., 2021) and Sentence Transformers clip-ViT-B-32-multilingual-v1 (Reimers and Gurevych, 2019)). A summary of results from these additional models is included in Table 7 and Table 8.

Model	English	Native Translated
NLLB SigLIP	31.3	28.5
Sentence Transformers MCLIP	24.9	18.3

Table 8: Average Recall over lower-income images across 25 countries for English and native translated language prompts for two multilingual LMM models.

5 Lessons Learned

We highlight key insights learned from our findings and present them below.

Current multilingual LMM models do not significantly improve diversity and representation. Our results from Section 4.1 demonstrate that English prompts perform better on lower (and higher) income images than prompts translated to a non-English language widely spoken in the region where the data was collected. Since the quality of translations, quantity of training data available for these languages, and consequently, the performance of AI models in these languages is lower than that of English, these findings are not very surprising. We can look forward to better non-English language performance as multilingual LMM models improve.

Location and socio-economic attributes improve retrieval performance for lower-income images. We find that adding geographical and socioeconomic attributes (including rich and neutral attributes) to prompts leads to an increased model preference for lower-income images over higher-income images, as demonstrated in Section 4.2. Images from poor households typically suffer the most from underrepresentation as they differ the most from the type of images available on the internet (Rosling et al., 2019). Since LMM models have learned representations from high-income images as the standard, then adding more information to the prompt (such as country suffixes like 'Malawi', income suffixes describing poverty or wealth, or neutral suffixes like 'a place') shifts the perspective of the model to retrieve images that are more diverse and less contained to the learned 'standard'.

Images with less standard topic appearances are retrieved using income suffix and country suffix prompts. Inspection of the retrieved images reveals that images with topic appearances commonly found in lower-income households previously not retrieved by the default English prompts are being retrieved with these prompts as shown in

Figure 1. For example, *pit latrines* and *forest-style toilets* previously left out by the default English prompts are retrieved using country suffixes (*Burundi* and *Cameroon*) and the *poor* income suffixes. Another example is "leaves" as "toilet paper" retrieved by *Liberia* and *Cameroon* country suffixes but excluded by the default English prompt.

6 Conclusion

In this paper, we addressed the uneven performance of LMM models across different countries and income levels. We explored three attribute-integrated prompting strategies: (1) translation of text prompts to native non-English languages, (2) addition of geographic information, and (3) addition of socioeconomic attributes. We found that integrating geographical and socioeconomic information into prompts enhances LMM model performance on images from lower-income households and retrieves more diverse label representations. Furthermore, we identified and highlighted the contexts where the proposed prompting techniques work best and shared our insights to improve representation in LMM models and datasets. Our code can be used to evaluate the performance of other LMM models and datasets and is publicly available at Analysis for Uplifting lower-income data.

Limitations

Translation Quality We note that, while NLLB-200-distilled-600M is reputed as a SOTA machine translation model, it does not have perfect accuracy on machine translation across all the languages it supports. We acknowledge that the quality of translations obtained from NLLB-200-distilled-600M greatly impacts our results.

Data Coverage Our study is constrained by the reach of the Dollar Street dataset and the number of contributions obtained from each region. Therefore, we do not account for data from other regions that are not included in the dataset.

Choice of Attributes We acknowledge that other attributes (e.g., physical attributes like color and material) of the objects in the images could be integrated into prompts to improve performance. However, we choose to focus on geographic and socioeconomic attributes since they are broad enough to include all possible topic appearances related to that attribute and their impact on data belonging to different countries and income groups can measured directly.

Diverse Data Availability While our methods facilitate the improvement of diversity during dataset annotation, these strategies cannot overcome the representation issues within the actual pool of images available for annotation.

Ethics Statement

Through this work, we aim to contribute toward improving diversity in AI models and even out the disparate impact of these models on the public, especially on underrepresented groups. The strategies discussed in our work can be used to prioritize the retrieval of lower-income images for balancing skewed data representation or domain-specific applications in AI. However, we do not encourage the use of these strategies to promote over-representation or the inclusion of one group over another in contexts that affect all members of the general public.

Our decision to use the NLLB-SigLIP model exemplifies our commitment to inclusive models that benefit as many people as possible, especially underrepresented groups. While researching technologically advanced communities is easier and less resource-intensive, we stress the importance of making AI design decisions that do not exclude communities with limited access to technology.

Acknowledgements

We are grateful to the Language and Information Technologies (LIT) lab members at the University of Michigan for their insightful discussions and feedback during the project's early stages. This project was partially funded by a grant from the Department of State (#STC10023GR0014), a grant from OpenAI and a Microsoft Foundational Model grant. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Department of State, OpenAI, or Microsoft.

References

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

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. 2022. Big vision. https://github.com/google-research/big_vision.

Kyle Buettner, Sina Malakouti, Xiang Lorraine Li, and Adriana Kovashka. 2024. Incorporating geo-diverse knowledge into prompting for increased geographical robustness in object recognition. *arXiv preprint arXiv:2401.01482*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Marcus Callies. 2024. Cultural conceptualisations in nigerian pidgin english proverbs. *World Englishes*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Emilio Ferrara. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3.

Emilio Ferrara. 2024. The butterfly effect in artificial intelligence systems: Implications for ai bias and fairness. *Machine Learning with Applications*, 15:100525.

Koustava Goswami, Srikrishna Karanam, Prateksha Udhayanan, Balaji Vasan Srinivasan, et al. 2023. Contextual prompt learning for vision-language understanding. *arXiv preprint arXiv:2307.00910*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. 2023. Diversity-aware meta visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10878–10887.

Oana Ignat, Longju Bai, Joan Nwatu, and Rada Mihalcea. 2024. Annotations on a budget: Leveraging geo-data similarity to balance model performance and annotation cost.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929.

Folgert Karsdorp and Lauren Fonteyn. 2019. Cultural entrenchment of folktales is encoded in language. *Palgrave Communications*, 5(1).

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. 2022. Prompt distribution learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5196–5205. IEEE Computer Society.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *arXiv preprint arXiv:1909.00871*.

Ercan G Nasif, Hamad Al-Daeaj, Bahman Ebrahimi, and Mary S Thibodeaux. 1991. Methodological problems in cross-cultural research: An updated review. *MIR: Management International Review*, pages 79–91.

Madeline Navarro, Camille Little, Genevera I Allen, and Santiago Segarra. 2024. Data augmentation via subgroup mixup for improving fairness. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7350–7354. IEEE.

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei Koh, and Ranjay Krishna. 2024. Multilingual diversity improves vision-language representations. *arXiv preprint arXiv:2405.16915*.

Bonny Norton. 1997. Language, identity, and the ownership of english. *TESOL quarterly*, 31(3):409–429.

Joan Nwatu, Oana Ignat, and Rada Mihalcea. 2023. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10686–10702, Singapore. Association for Computational Linguistics.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversityin contrastive vision-language models. *arXiv preprint arXiv:2405.13777*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481.

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. *arXiv preprint arXiv:2301.02560*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *Advances in Neural Information Processing Systems*, 35:12979–12990.

Han Rosling, Ola Rosling, and Anna Rosling Rönnlund. 2019. *Factfulness*. Lindhardt og Ringhof.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.

Farzad Sharifian. 2014. Language and culture: Overview. *The Routledge handbook of language and culture*, pages 3–17.

Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 358–364, New York, NY, USA. Association for Computing Machinery.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Ashish Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *EMNLP*.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2023. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *arXiv preprint arXiv:2310.01929*.

Alexander Visheratin. 2023. Nllb-clip–train performant multilingual image retrieval model on a budget. *arXiv preprint arXiv:2309.01859*.

Robert F Woolson. 2005. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics*, 8.

Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *International Conference on Machine Learning*, pages 42602–42613. PMLR.

A Appendix

A.1 Non-English Languages

We use the following non-English languages in our experiments. *'German', 'Spanish', 'Portuguese', 'French', 'Chinese', 'Czech', 'Danish', 'Arabic', 'Hindi', 'Indonesian', 'Farsi-Persian', 'Italian', 'Russian', 'Mongolian', 'Burmese', 'Dutch', 'Urdu', 'Romanian', 'Serbian', 'Korean', 'Swedish', 'Thai', 'Turkish', 'Ukrainian', 'Vietnamese', 'Bengali', 'Khmer', 'Oromo', 'Ewe', 'Creole', 'Swahili', 'Nepali', 'Hausa', 'Kyrgyz', 'Tagalog', 'Kinyarwanda', 'Somali', 'Zulu', 'Sinhala', 'Shona'*

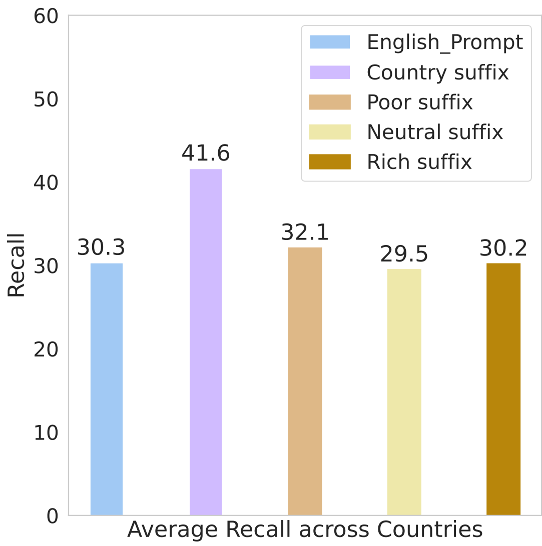

Figure 6: Average Recall over lower-income images across 39 countries for English, Country Suffix, and Income Suffix prompts

Countries	Non-English Language	Image Income Classes	World Bank Country Economic Classes	Continent
Austria	German	Rich, Up-mid	High	Europe
Bangladesh	Bengali	Poor	Low-mid	Asia
Bolivia	Spanish	Low-mid, poor	Low-mid	America
Brazil	Portuguese	Rich, Up-mid, Low-mid	Up-mid	America
Burkina Faso	French	Poor	Poor/low	Africa
Burundi	French	Poor	Poor/low	Africa
Cambodia	Khmer	Up-mid, Low-mid, Poor	Low-mid	Asia
Cameroon	French	Up-mid, Low-mid, Poor	Low-mid	Africa
Canada	French	Rich	High	America
China	Chinese	Rich, Up-mid, Low-mid, Poor	Up-mid	Asia
Colombia	Spanish	Rich, Up-mid, Low-mid, Poor	Up-mid	America
Cote d'Ivoire	French	Poor	Low-mid	Africa
Czech Republic	Czech	Rich	High	Europe
Denmark	Danish	Rich	High	Europe
Egypt	Arabic	Up-mid	Low-mid	Africa
Ethiopia	Oromo	Rich, Up-mid, Low-mid	Poor/low	Africa
France	French	Rich, Up-mid	High	Europe
Ghana	Ewe	Low-mid	Low-mid	Africa
Guatemala	Spanish	Low-mid	Up-mid	America
Haiti	Creole	Poor	Low-mid	America
India	Hindi	Rich, Up-mid, Low-mid, Poor	Low-mid	Asia
Indonesia	Bahasa Indonesian	Rich, Up-mid, Low-mid, Poor	Up-mid	Asia
Iran	Farsi (Persian)	Rich, Up-mid	Low-mid	Asia
Italy	Italian	Rich	High	Europe
Jordan	Arabic	Rich, Low-mid	Low-mid	Asia
Kazakhstan	Russian	Up-mid	Up-mid	Asia
Kenya	Swahili	Rich, Low-mid, Poor	Low-mid	Africa
Kyrgyzstan	Kyrgyz	Up-mid	Low-mid	Asia
Lebanon	Arabic	Up-mid	Low-mid	Asia
Liberia	-	Poor	Poor/low	Africa
Malawi	-	Poor	Poor/low	Africa
Mexico	Spanish	Rich, Up-mid	Up-mid	America
Mongolia	Mongolian	Low-mid	Low-mid	Asia
Myanmar	Burmese	Low-mid, poor	Low-mid	Asia
Nepal	Nepali	Rich, Up-mid, Low-mid, Poor	Low-mid	Asia
Netherlands	Dutch	Rich, Up-mid	High	Europe
Nigeria	Hausa	Rich, Up-mid, Low-mid, Poor	Low-mid	Africa
Pakistan	Urdu	Rich, Up-mid, Low-mid, Poor	Low-mid	Asia
Palenstine	Arabic	Low-mid, poor	Low-mid	Asia
Papua New Guinea	-	Poor	Low-mid	Asia
Peru	Spanish	Low-mid, poor	Up-mid	America
Philippines	Tagalog	Up-mid, Low-mid, Poor	Low-mid	Asia
Romania	Romanian	Rich	High	Europe
Russia	Russian	Rich, Up-mid	Up-mid	Europe
Rwanda	Kinyarwanda	Low-mid, poor	Poor/low	Africa
Serbia	Serbian	Rich, Up-mid, Low-mid	Up-mid	Europe
Somalia	Somali	Poor	Poor/low	Africa
South Africa	Zulu	Rich, Up-mid, Low-mid, Poor	Up-mid	Africa

Countries	Non-English Language	Image Income Classes	World Bank Country Economic Classes	Continent
South Korea	Korean	Rich, Up-mid, Low-mid	High	Asia
Spain	Spanish	Rich	High	Europe
Sri Lanka	Sinhala	Up-mid	Low-mid	Asia
Sweden	Swedish	Rich, Up-mid	High	Europe
Switzerland	German	Rich	High	Europe
Tanzania	Swahili	Up-mid, Low-mid, Poor	Low-mid	Africa
Thailand	Thai	Up-mid, Low-mid, Poor	Up-mid	Asia
Togo	French	Low-mid, poor	Poor/low	Africa
Tunisia	Arabic	Low-mid, poor	Low-mid	Africa
Turkey	Turkish	Rich	Up-mid	Europe
Ukraine	Ukrainian	Rich, Up-mid, Low-mid	Low-mid	Europe
United Kingdom	-	Rich, Up-mid	High	Europe
United States	Spanish	Rich, Up-mid, Low-mid	High	America
Vietnam	Vietnamese	Low-mid, Rich	Low-mid	Asia
Zimbabwe	Shona	Poor	Low-mid	Africa

Table 9: Table displaying the 63 Dollar Street countries, their major non-English language, income levels of contributions for that country, World Bank income class, and their continent.

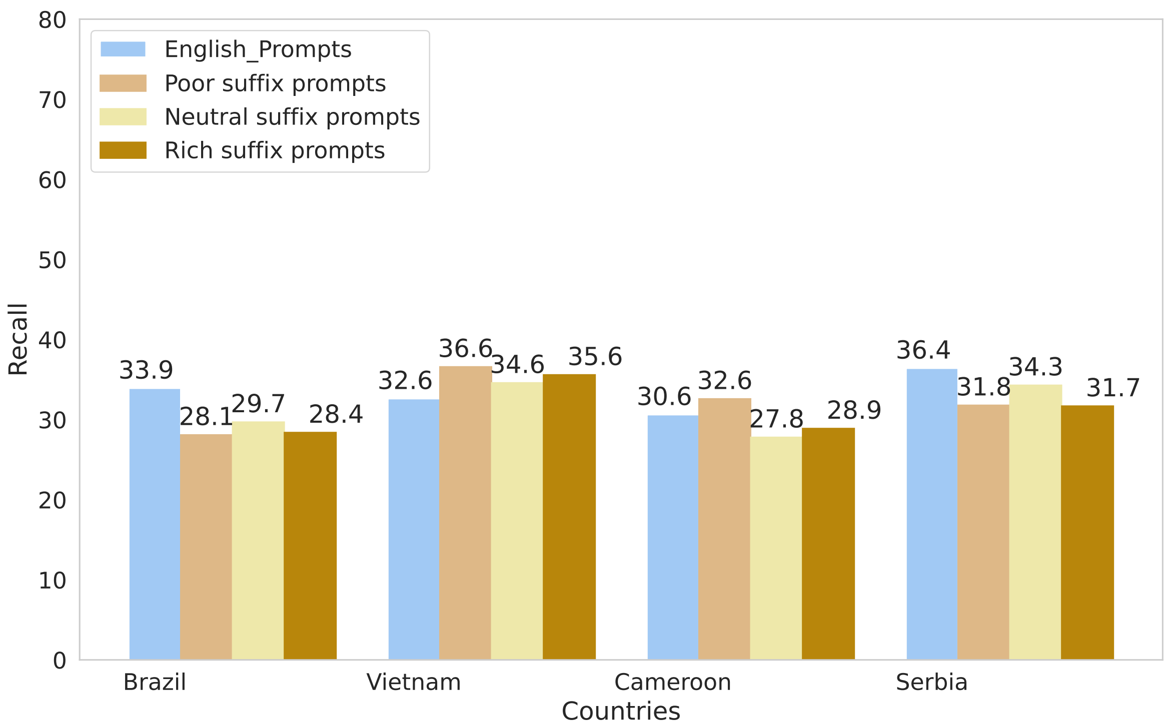

Figure 7: Average NLLB SigLIP Recall over poor and lower-middle income images for English and Income Suffix prompts

Languages	Income level of images			
	Poor	Low-mid	Up-mid	Rich
Arabic	21.0 (-2.4)	32.0 (-3.9)	38.5 (-7.1)	37.3 (-6.3)
Bengali	20.9 (-1.5)	33.0 (-2.9)	40.7 (-4.9)	38.9 (-4.7)
Burmese	21.5 (-0.9)	30.4 (-5.5)	36.0 (-9.6)	34.2 (-9.4)
Chinese	21.5 (-0.9)	32.5 (-3.4)	39.1 (-6.5)	37.3 (-6.3)
Creole	21.0 (-1.4)	32.6 (-3.3)	40.1 (-5.5)	38.0 (-5.6)
Czech	19.9 (-2.5)	32.3 (-3.6)	40.3 (-5.3)	38.8 (-4.8)
Danish	20.8 (-1.6)	33.7 (-2.2)	41.7 (-3.9)	40.0 (-3.6)
Dutch	21.1 (-1.3)	33.6 (-2.3)	42.5 (-3.1)	40.8 (-2.8)
Ewe	14.7 (-7.7)	19.3 (-16.6)	22.1 (-23.5)	20.6 (-23.0)
Farsi-Persian	21.9 (-0.5)	31.8 (4.1)	39.1 (-6.5)	38.1 (-5.5)
French	21.7 (-0.7)	33.7 (-2.2)	42.6 (-3.0)	41.4 (-2.2)
German	21.5 (-0.9)	33.1 (-2.8)	41 (-4.6)	39 (-4.6)
Hausa	20.6 (-1.8)	31.6 (-4.3)	38.4 (-7.2)	36.4 (-7.2)
Hindi	22.2 (-0.2)	34.5 (-1.4)	41.8 (-3.8)	40 (-3.6)
Indonesian	22.1 (-0.3)	34.8 (-1.1)	42.4 (-3.2)	40.5 (-3.1)
Italian	21.1 (-1.3)	34.3 (-1.6)	42.7 (-2.9)	41.4 (-2.2)
Khmer	16.4 (-6.0)	22.0 (-13.9)	24.8 (-20.8)	23.2 (-20.4)
Kinyarwanda	16.7 (-5.7)	23.7 (-12.2)	28.8 (-16.8)	27.2 (-16.4)
Korean	20.4 (-2.0)	32.2 (-3.7)	40.2 (-5.4)	37.9 (-5.7)
Kyrgyz	21.6 (-0.8)	30.9 (-5.0)	36.7 (-8.9)	35.7 (-7.9)
Mongolian	13.7 (-8.7)	20.9 (-1.5)	25 (-20.6)	23.4 (20.2)
Nepali	20.7 (-1.7)	32.5 (-3.4)	40.9 (-4.7)	39.7 (-3.9)
Oromo	15.8 (-6.6)	20.9 (-15.0)	24.7 (-20.9)	23.4 (20.2)
Portuguese	21.3 (-1.1)	34 (-1.9)	42.6 (-3.0)	41.2 (-2.4)
Romanian	20.3 (-2.1)	32.9 (-3.0)	41.0 (-4.6)	38.9 (-4.7)
Russian	21.1 (-1.3)	33.4 (-2.5)	41.5 (-4.1)	39.9 (-3.7)
Serbian	19.2 (-3.2)	30.8 (-5.1)	37.2 (-8.4)	35.6 (-8.0)
Shona	19.1 (-3.3)	27.2 (-8.7)	32.2 (-13.4)	30.5 (-13.1)
Sinhala	20.4 (-2.0)	32.0 (-3.9)	37.9 (-7.7)	35.7 (-7.9)
Somali	19.0 (-3.4)	28.5 (-7.4)	33.8 (-11.8)	31.4 (-12.2)
Spanish	20.7 (-1.7)	33.8 (-2.1)	42.5 (-3.1)	40.9 (-2.7)
Swahili	22.1 (-0.3)	33.6 (-2.3)	41.3 (-4.3)	38.9 (-4.7)
Swedish	20.5 (-1.9)	33.0 (-2.9)	40.5 (-5.1)	38.6 (-5.0)
Tagalog	21.4 (-1.0)	33.2 (-2.7)	39.4 (-6.2)	37.5 (-6.1)
Thai	19.7 (-2.7)	29.7 (-6.2)	34.9 (-10.7)	33.6 (-10.0)
Turkish	20.5 (-1.9)	31.6 (-4.3)	39.5 (-6.1)	38.5 (-5.1)
Ukrainian	20.7 (-1.7)	33.0 (-2.9)	40.7 (-4.9)	38.7 (-4.9)
Urdu	21.5 (-0.9)	33.0 (-2.9)	40.6 (-5.0)	39.1 (-4.5)
Vietnamese	20.6 (-1.8)	32.8 (-3.1)	41.1 (-4.5)	39.5 (-4.1)
Zulu	19.9 (-2.5)	29.9 (-6.0)	35.2 (-10.4)	33.4 (-10.2)

Table 10: Non-English prompts lead to a decrease in Recall scores across all income levels. Table of the differences (rounded to 1 d.p.) between Recall scores for non-English language prompts and Recall scores for default English prompts for all data grouped into income levels.

Country Suffix	Income levels			
	Poor Δ	Low-mid Δ	Up-mid Δ	Rich Δ
Burkina Faso	**0.327** (+0.103)	0.303 (-0.056)	0.227 (-0.229)	0.185 (-0.251)
Burundi	**0.331** (+0.107)	0.279 (-0.08)	0.197 (-0.259)	0.154 (-0.282)
Ethiopia	**0.334** (+0.11)	0.313 (-0.046)	0.269 (-0.187)	0.227 (-0.209)
Liberia	**0.327** (+0.103)	0.303 (-0.056)	0.249 (-0.207)	0.205 (-0.231)
Malawi	0.301 (+0.077)	**0.32** (-0.39)	0.286 (-0.17)	0.254 (-0.182)
Rwanda	**0.334** (+0.11)	0.322 (-0.037)	0.249 (-0.207)	0.204 (-0.232)
Somalia	**0.318** (+0.094)	0.296 (-0.063)	0.243 (-0.213)	0.209 (-0.227)
Togo	0.297 (+0.073)	**0.31** (-0.049)	0.283 (-0.173)	0.252 (-0.184)
Bangladesh	0.271 (+0.047)	**0.319** (-0.04)	0.266 (-0.19)	0.221 (-0.215)
Bolivia	0.3 (+0.076)	**0.318** (-0.041)	0.262 (-0.194)	0.223 (-0.213)
Cambodia	**0.289** (+0.065)	0.288 (-0.071)	0.213 (-0.243)	0.172 (-0.264)
Cameroon	**0.313** (+0.089)	0.289 (-0.07)	0.233 (-0.223)	0.192 (-0.244)
Cote d'Ivoire	0.23 (+0.006)	0.296 (-0.063)	**0.325** (-0.131)	0.302 (-0.134)
Egypt	0.257 (+0.033)	0.334 (-0.025)	**0.357** (-0.099)	0.316 (-0.12)
Ghana	**0.314** (+0.09)	0.294 (-0.065)	0.267 (-0.189)	0.233 (-0.203)
Haiti	0.296 (+0.072)	**0.331** (-0.028)	0.307 (-0.149)	0.269 (-0.167)
India	0.239 (+0.015)	**0.31** (-0.049)	0.306 (-0.15)	0.278 (-0.158)
Iran	0.221 (-0.003)	0.343 (-0.016)	**0.375** (-0.081)	0.337 (-0.099)
Jordan	0.222 (-0.002)	0.308 (-0.051)	**0.376** (-0.08)	0.371 (-0.065)
Kenya	0.296 (+0.072)	**0.318** (-0.041)	0.283 (-0.173)	0.236 (-0.2)
Kyrgyzstan	0.229 (+0.005)	0.338 (-0.021)	**0.365** (-0.091)	0.318 (-0.118)
Lebanon	0.249 (+0.025)	0.309 (-0.05)	**0.348** (-0.108)	0.33 (-0.106)
Mongolia	0.259 (+0.035)	**0.326** (-0.033)	0.308 (-0.148)	0.256 (-0.18)
Myanmar	0.263 (+0.039)	**0.304** (-0.055)	0.241 (-0.215)	0.195 (-0.241)
Nepal	0.274 (+0.05)	**0.307** (-0.052)	0.253 (-0.203)	0.213 (-0.223)
Nigeria	**0.294** (+0.07)	0.286 (-0.073)	0.256 (-0.2)	0.223 (-0.213)
Pakistan	0.197 (-0.027)	0.303 (-0.056)	**0.321** (-0.135)	0.289 (-0.147)
Palestine	0.258 (+0.034)	0.349 (-0.01)	**0.361** (-0.095)	0.317 (-0.119)
Papua New Guinea	0.274 (+0.05)	**0.302** (-0.057)	0.266 (-0.19)	0.235 (-0.201)
Philippines	0.271 (+0.047)	**0.346** (-0.013)	0.337 (-0.119)	0.295 (-0.141)
Sri Lanka	0.275 (+0.051)	**0.322** (-0.037)	0.303 (-0.153)	0.277 (-0.159)
Tanzania	0.287 (+0.063)	**0.292** (-0.067)	0.257 (-0.199)	0.228 (-0.208)
Tunisia	0.276 (+0.052)	**0.321** (-0.038)	0.314 (-0.142)	0.284 (-0.152)
Ukraine	0.245 (+0.021)	0.355 (-0.004)	**0.372** (-0.084)	0.323 (-0.113)
Vietnam	0.229 (+0.005)	0.321 (-0.038)	**0.33** (-0.126)	0.294 (-0.142)
Zimbabwe	**0.312** (+0.088)	0.311 (-0.048)	0.285 (-0.171)	0.242 (-0.194)

Table 11: Table of low-income/poor (in lilac) and lower-middle income (in purple) country suffixes and their effect on Recall for different income groups. For each country suffix, the highest Recall among income groups is highlighted in bold. The green and red values show how much increase or reduction that country suffix has on the Recall of data from an income group compared to default English prompts.

Country Suffix	Income levels			
	Poor Δ	**Low-mid** Δ	**Up-mid** Δ	**Rich** Δ
Brazil	0.254 (+0.03)	0.303 (-0.056)	**0.323** (-0.133)	0.303 (-0.133)
China	0.213 (-0.011)	0.34 (-0.019)	**0.369** (-0.087)	0.319 (-0.117)
Colombia	0.3 (+0.076)	**0.324** (-0.035)	0.275 (-0.181)	0.232 (-0.204)
Guatemala	0.269 (+0.045)	**0.314** (-0.045)	0.277 (-0.179)	0.233 (-0.203)
Indonesia	0.266 (+0.042)	**0.328** (-0.031)	0.303 (-0.153)	0.266 (-0.17)
Kazakhstan	0.254 (+0.03)	**0.337 (-0.022)**	**0.337** (-0.119)	0.292 (-0.144)
Mexico	0.251 (+0.027)	0.335 (-0.024)	**0.357** (-0.099)	0.312 (-0.124)
Peru	0.261 (+0.037)	**0.319** (-0.04)	0.317 (-0.139)	0.287 (-0.149)
Russia	0.212 (-0.012)	0.344 (-0.015)	**0.382** (-0.074)	0.335 (-0.101)
Serbia	0.197 (-0.027)	0.313 (-0.046)	**0.378** (-0.078)	0.354 (-0.082)
South Africa	0.291 (+0.067)	**0.302** (-0.057)	**0.302** (-0.154)	0.269 (-0.167)
Thailand	0.234 (+0.01)	**0.312** (-0.047)	0.293 (-0.163)	0.256 (-0.18)
Turkey	0.228 (+0.004)	0.321 (-0.038)	**0.333** (-0.123)	0.302 (-0.134)
Austria	0.166 (-0.058)	0.296 (-0.063)	0.407 (-0.049)	**0.408** (-0.028)
Canada	0.266 (+0.042)	0.355 (-0.004)	**0.391** (-0.065)	0.359 (-0.077)
Czech Republic	0.195 (-0.029)	0.33 (-0.029)	**0.395** (-0.061)	0.379 (-0.057)
Denmark	0.184 (-0.04)	0.293 (-0.066)	0.386 (-0.07)	**0.394** (-0.042)
France	0.199 (-0.025)	0.317 (-0.042)	0.415 (-0.041)	**0.417** (-0.019)
Italy	0.192 (-0.032)	0.318 (-0.041)	**0.379** (-0.077)	0.367 (-0.069)
Netherlands	0.219 (-0.005)	0.31 (-0.049)	**0.374** (-0.082)	0.359 (-0.077)
Romania	0.255 (+0.031)	0.337 (-0.022)	**0.342** (-0.114)	0.312 (-0.124)
South Korea	0.225 (+0.001)	0.313 (-0.046)	**0.345** (-0.111)	0.314 (-0.122)
Spain	0.183 (-0.041)	0.308 (-0.051)	**0.413** (-0.043)	0.404 (-0.032)
Sweden	0.167 (-0.057)	0.294 (-0.065)	0.405 (-0.051)	**0.412** (-0.024)
Switzerland	0.135 (-0.089)	0.257 (-0.102)	0.363 (-0.093)	**0.389** (-0.047)
United Kingdom	0.205 (-0.019)	0.326 (-0.033)	**0.418** (-0.038)	0.409 (-0.027)
United States	0.25 (+0.026)	0.362 (-0.003)	**0.421** (-0.035)	0.391 (0.045)

Table 12: Table of high-income/rich (in blue) and upper-middle-income (in sky blue) country suffixes and their effect on Recall for different income groups. For each country suffix, the highest Recall among income groups is highlighted in bold. The green and red values show how much increase or reduction that country suffix has on the Recall of data from an income group compared to default English prompts.

Language	ISO	chrf++
Arabic	arb_Arab	51.4
Bengali	ben_Beng	46.2
Burmese	mya_Mymr	29.3
Chinese	zho_Hans	19.6
Creole	hat_Latn	50.2
Czech	ces_Latn	52.7
Danish	dan_Latn	63.2
Dutch	nld_Latn	53.1
Ewe	ewe_Latn	35.6
Farsi_Persian	pes_Arab	47.4
French	fra_Latn	67.0
German	deu_Latn	59.4
Hausa	hau_Latn	49.0
Hindi	hin_Latn	54.2
Indonesian	ind_Latn	66.6
Italian	ita_Latn	54.6
Khmer	khm_Khmr	31.2
Kinyarwanda	kin_Latn	44.0
Korean	kor_Hang	32.1
Kyrgyz	kir_Cyrl	42.6
Mongolian	khk_Cyrl	37.3
Nepali	npi_Deva	49.0
Oromo	gaz_Latn	31.6
Portuguese	por_Latn	67.4
Romanian	ron_Latn	58.2
Russian	rus_Cyrl	52.5
Serbian	srp_Cyrl	53.3
Shona	sna_Latn	42.9
Sinhala	sin_Sinh	42.4
Somali	som_Latn	41.5
Spanish	spa_Latn	52.6
Swahili	swh_Latn	58.0
Swedish	swe_Latn	62.7
Tagalog	tgl_Latn	56.4
Thai	tha_Thai	36.0
Turkish	tur_Latn	52.9
Ukrainian	ukr_Cyrl	50.5
Urdu	urd_Arab	46.6
Vietnamese	vie_Latn	56.4
Zulu	zul_Latn	51.0

Table 13: Languages used and translation metrics (chrf++ scores) for NLLB-200-distilled-600M from English to these languages.

Prompt	P-value	Sig. or not
English & Native translated	8.64e-09	yes
English & Country suffix	7.27e-08	yes
English & Poor Income Suffix	0.02	yes
English & Rich Income Suffix	0.603	no
English & Neutral Income Suffix	0.563	no

Table 14: Table showing p-values of Wilcoxon test between the default English prompt and each of the formulated prompts. The difference is regarded as statistically significant when $p \leq 0.05$.