The Echoes of Multilinguality:
Tracing Cultural Value Shifts during LM Fine-tuning

Rochelle Choenni
University of Amsterdam
r.m.v.k.choenni@uva.nl

Anne Lauscher
University of Hamburg
anne.lauscher@uni-hamburg.de

Ekaterina Shutova
University of Amsterdam
e.shutova@uva.nl

Abstract

Texts written in different languages reflect different culturally-dependent beliefs of their writers. Thus, we expect multilingual LMs (MLMs), that are jointly trained on a concatenation of text in multiple languages, to encode different cultural values for each language. Yet, as the 'multilinguality' of these LMs is driven by cross-lingual sharing, we also have reason to belief that cultural values bleed over from one language into another. This limits the use of MLMs in practice, as apart from being proficient in generating text in multiple languages, creating language technology that can serve a community also requires the output of LMs to be sensitive to their biases (Naous et al., 2023). Yet, little is known about how cultural values emerge and evolve in MLMs (Hershcovich et al., 2022a). We are the first to study how languages can exert influence on the cultural values encoded for different test languages, by studying how such values are revised during fine-tuning. Focusing on the fine-tuning stage allows us to study the interplay between value shifts when exposed to new linguistic experience from different data sources and languages. Lastly, we use a training data attribution method to find patterns in the fine-tuning examples, and the languages that they come from, that tend to instigate value shifts.

1 Introduction

Training LMs on large text corpora has been shown to induce various types of (social) biases in multilingual LMs (MLMs) that affect which human values the model picks up on (Choenni et al., 2021; Hämmerl et al., 2022). However, human values vary per culture, which means that the cultural values that are reflected through their language (either explicitly or implicitly) will also differ. As MLMs are trained on the concatenation of text from a wide variety of languages spoken in the world, we can expect different, and perhaps opposing, cultural values to be encoded in them simultaneously. This necessitates MLMs to become inherently multicultural as well in order to appropriately serve culturally diverse communities (Naous et al., 2023; Talat et al., 2022). In fact, it has already been shown that MLMs encode a distinct set of cultural values for different languages. However, those values do not tend to align with those collected from real human surveys conducted in the countries where the majority population speak the respective languages (Arora et al., 2023; Kovač et al., 2023). As such, the multilingual NLP community is now faced with the new pressing challenge of better culturally aligning MLMs to human values (Yao et al., 2023). Thus, we aim to study how cultural values emerge and evolve in MLMs to better understand and aid cross-cultural value alignment in the future.

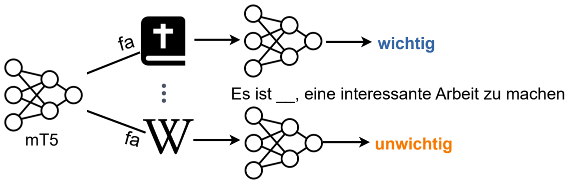


Figure 1: An example of our evaluation setup. We analyze the effect that fine-tuning on different data sources in a fine-tuning language A (Farsi) has on the cultural values that are encoded for a test language B (German). Translation of the example: *'It is _ to do interesting work.'* (options: *'important'* or *'unimportant'*).

In particular, we hypothesise that training on multilingual data leads to an interaction of language-specific cultural values within the models, possibly steering a language's cultural bias into a direction that is unfaithful to the majority of that language's speakers. This raises an interesting set of questions on how languages exert influence on the encoding of cultural values. Focusing on the fine-tuning stage, we study how cultural values are revised during training. For instance, given a set of fine-tuning languages, test languages, and data sources, when we fine-tune on a data source in a

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15042–15058
August 11-16, 2024 ©2024 Association for Computational Linguistics

language A, and test in a language B, do we inadvertently induce the cultural values from A into B? And would the same effect be visible across all test languages or are the values encoded for some languages more prone to change? Moreover, how much impact does the bias of the data source itself have versus the language used for fine-tuning? And can different data sources systematically enforce different alignments to human values?

To better understand this cross-cultural interaction, we study the following questions: (**Q1**) How do the fine-tuning language and data source affect the way in which cultural information is encoded and revised during fine-tuning? (**Q2**) How do cultural value shifts change the alignment to human survey data? (**Q3**) Can we find patterns in the training examples that systematically influence how cultural values are revised? To this end, we conduct a set of controlled fine-tuning experiments using multi-parallel texts from data sources with neutral bias (Wikipedia), religious bias (Bible and Quran) and political bias (news articles) across 4 fine-tuning languages and 13 test languages. We follow Arora et al. (2023) in using 200+ WVS survey questions to probe for cross-cultural values in pretrained and fine-tuned MLMs. Importantly, using survey questions as probes allows us to test the alignment between model predictions and human data. Finally, we use a training data attribution (TDA) to trace value shifts back to the data source.

We find that, while fine-tuning language and domain source play a minor role in the revision of cultural information compared to the amount of fine-tuning data, fine-tuning languages can lead the cultural perspective of test languages into different directions. Importantly, this can positively affect the models' alignment to human values. Yet, overall, results vary considerably across test languages. Moreover, our TDA analysis provides interesting insights about the systematicity with which the model tends to rely on parallel data to instigate the same value shifts across languages. Our results underpin the complexity of cross-language and cross-cultural interaction within MLMs, and suggest that the semantic content of fine-tuning data might not be the main reason for value shifts.

2 Related work

2.1 Cross-cultural NLP

The fact that LMs are becoming increasingly multilingual, has given rise to a new subfield in NLP that is concerned with questions such as to what extent these models are multicultural (Liu et al., 2023; Havaldar et al., 2023; Hershcovich et al., 2022a), to what extent the cultural values that they encode align with those from human populations (Naous et al., 2023; Arora et al., 2023; Cao et al., 2023), and whether we can automatically improve such an alignment to better serve culturally diverse communities (Kovač et al., 2023). For instance, Naous et al. (2023) show that MLMs tend to exhibit western-centric biases, even when being prompted in Arabic and contextualized by an Arab cultural setting, resulting in culturally insensitive output such as suggesting to go for a beer after Islamic prayer. Similarly, previous works show that LMs fail to understand proverbs and sayings from different languages (Liu et al., 2023), and do not capture the nuances in meaning and usage patterns of emotion words that exist differently across cultures (Havaldar et al., 2023). These findings suggest that there is still an important gap to fill when it comes to creating multilingual language technology that is also multicultural (Talat et al., 2022). We aim to contribute to our understanding of how cultural values manifest across languages.

2.2 Probing for bias

Cloze-style testing is a technique that stems from psycholinguistics (Taylor, 1953), and has been popularized as a tool to study different types of knowledge and biases encoded by LMs. The idea is that we prompt LMs with a carefully curated set of probing sentences that are meant to elicit responses that expose the biases encoded within the LM (May et al., 2019; Stańczak et al., 2023; Nangia et al., 2020). While many different types of biases have been studied in the multilingual setting (Hämmerl et al., 2022; Touileb et al., 2022; Kaneko et al., 2022), Arora et al. (2023) are the first to probe for cross-cultural values in pretrained MLMs. We use their probing questions in a similar set-up, but take a step further by studying how different fine-tuning languages can exert influence on cultural values encoded for a different set of test languages.

2.3 Training data attribution

Training data attribution (TDA) methods are developed to identify a set of training examples that were most influential in making a particular test prediction. The influence of a training example z_{train} on test example z_{test} can typically be formalized as the change in loss that would have been incurred for z_{test}, if sample z_{train} was not seen during training (Koh and Liang, 2017). Thus, we can use the resulting influence scores as a measure of how important z_{train} is for making a prediction for z_{test}. TDA methods have successfully been used on classification tasks in NLP (Han and Tsvetkov, 2022), e.g., to detect outlier data (Han et al., 2020; Lam et al., 2022) or to correct model predictions (Meng et al., 2020; Guo et al., 2021). Recently, they have been applied to study cross-lingual sharing in MLMs (Choenni et al., 2023a,b). Yet, extending the use of TDA methods beyond classification tasks has proved difficult. Akyürek et al. (2022) first used TDA methods (Rajani et al., 2019; Pruthi et al., 2020) on masked language modelling for fact tracing – the task of attributing an LM's factual assertions back to training examples. Yet, the results were shown to be unreliable. More recently, however, Park et al. (2023) proposed TRAK, which was shown to be successful in *behaviour tracing* on mT5. We adopt their approach to trace mT5 predictions for cloze-style questions eliciting cultural values back to the fine-tuning data.

3 Methodology

3.1 World Values Survey (WVS)

We probe for cultural values using cloze-style testing templates derived from the questions proposed in the World Values Survey (WVS) (Haerpfer et al., 2022) by Arora et al. (2023). Thus, more precisely, we study descriptive ethics (Vida et al., 2023). The WVS collects data on cultural values in different countries in waves, and our questions come from Wave 7 which ran from 2017 to 2020 and targets 57 countries [1]. Survey results are published per question, organised in 13 categories: (1) Social Values, Attitudes and Stereotypes, (2) Happiness and Well-being, (3) Social Capital, Trust and Organisational Membership, (4) Economic Values, (5) Corruption, (6) Migration, (7) Security, (8) Postmaterialist Index, (9) Science and Technology, (10) Religious Values, (11) Ethical Values and Norms, (12) Political Interest and Political Participation, (13) Political Culture and Regimes. Categories (4) and (8) are excluded as their questions could not be converted into probes. We use 237 probes in total.

3.2 Multilingual probes

We use the English probes that were professionally translated into the following 13 languages: Romanian, Greek, Urdu, Farsi, Tagalog, Indonesian, German, Malay, Bengali, Serbian, Turkish, Vietnamese and Korean, see Figure 1 for an example. Note that these languages were carefully selected by Arora et al. (2023) based on the following three criteria: (1) the languages can be mapped to one country covered by the WVS survey, (2) the languages are the official languages of the countries that they are mapped to, (3) the distribution of the language's speakers can be primarily localized to a country or relatively small geographical region, and (4) all selected languages have at least 10K articles on Wikipedia such that the LMs have seen a sufficient amount of pretraining data.

3.3 Models

Arora et al. (2023) report that cultural information is inconsistent across different pretrained LMs. Given recent trends on scaling LMs to tens of billions of parameters (Scao et al., 2022), we study how model size affects cultural information instead. We probe mT5 (Xue et al., 2021) small, base and large that contain 0.3B, 0.58B and 1.2B parameters.

3.4 Probing method

To probe for cultural values, we query the mT5 models with the cloze-style question probes from Section 3.2 using a conditional language generation head. More concretely, for each probing template, we replace the `[MASK]` token in the original probes with extra ID tokens for mT5, and apply softmax over the logits of all tokens in the vocabulary V. We then take the log probability for the two candidate answers of the question, and take the option with the highest log probability as the final answer.

3.5 Quantifying shifts in cultural profiles

To compare overall cultural bias across languages and model sizes, we build 'cultural profiles' based on their predictions for all WVS questions. Per question, we take the log probabilties of the respective answers and apply softmax to them to obtain the probabilities for selecting the first answer option for all N questions. We then compile them into a N-dimensional vector, which represents the cultural profile of a given language. Similarly, we obtain ground truth profiles for the corresponding countries using the results from the WVS survey. The results are reported as the percentage of interviewees that selected each class. Yet, in contrast to our probes, the survey proposes multiple classes

[1] https://www.worldvaluessurvey.org

	bn	de	el	fa	id	ko	ms	ro	sr	tl	tr	ur	vi	avg.
S vs. B	78	89	84	89	77	82	84	94	72	70	79	87	83	82
S vs. L	78	87	83	89	69	86	81	96	63	69	75	83	83	80
B vs. L	88	86	81	87	79	83	84	94	72	77	81	85	87	83

Table 1: Percentage of agreement between model predictions from the pretrained mT5 **S**mall, **B**ase and **L**arge models fine-tuned on PBC (10K).

(e.g. *'very important'*, *'important'*, *'not very important'*, *'not important'*). We add the probability from the middle classes to the closest class on either end of the spectrum e.g., very important/important becomes one class. We then test how similar cultural profiles within pretrained models are to the ground truth, and in which direction they change after fine-tuning, by computing the change in correlation.

4 Experimental setup

4.1 Data sources

We use three data sources with multi-parallel data for fine-tuning: Flores-101 (Goyal et al., 2022), the Parallel Bible Corpus (PBC) (Christodouloupoulos and Steedman, 2015), and the Tanzil dataset (Tiedemann, 2012) that contain human translated sentences from Wikipedia articles, Bible texts and Quran texts respectively. While Flores-101 is more likely to be used in practice, PBC and Tanzil are an interesting testbed as due to their didactic nature, we expect cultural values to be affected more heavily. We select 4 languages for fine-tuning: Farsi, Korean, Hindi and Russian. These languages all rely on a different writing script, and are commonly spoken by culturally diverse populations. Also, Farsi and Korean are included in our test languages. The PBC dataset already contains multi-parallel sentences, and for the Tanzil we were able to extract them automatically using the English sentences in the translation pairs. Finally, following Choenni et al. (2021), we also fine-tune on articles from different news sources across the political spectrum from left to right-wing ideologies. We use English news articles collected between 2013 and early 2020 from New Yorker (*left*), Reuters (*center*) and FOX news (*right*) from the *All-The-News* dataset. While the focus of this study is on language influence, we use this as an additional test to disentangle the effect of language bias from domain bias.

4.2 Training details

From each data source we use either 2K or 10K consecutive sentences for fine-tuning on the MLM 'span corruption' objective that was used for pretraining, see Appendix B for training details. We use two fine-tuning strategies (FT): (1) monolingual FT, where we train our models on each language separately, and (2) multilingual FT, where we jointly train on all fine-tuning languages together. For (2), we use 2.5K multi-parallel sentences for each language and shuffle them before training. We compare multilingual and monolingual models where 10K sentences are seen in total.

5 Probing results

As a baseline to our fine-tuning experiments, we first study the cultural profiles encoded in the pretrained models. In Section 5.2, we then analyze how cultural information in the model changes as a result of cross-language and domain influence.

5.1 Cultural information in pretrained LMs

As explained in Section 3.5, we build cultural profiles for each country and compute Spearman correlation between the ground truth and pretrained model profiles. In line with previous results (Arora et al., 2023), we confirm that all pretrained LMs correlate poorly with human values. Yet, in Table 1, we find that the models of varying sizes on average agree on 80% of the survey questions (pairwise). In addition, we find that variations in consistency mostly depend on the test language. For instance, in Romanian the models agree on $\geq 94\%$ of questions, but for Serbian this is $\leq 72\%$ instead. Similarly, averaged across test languages, the models agree more on specific WVS categories e.g., predictions are more consistent for questions pertaining to happiness, security and political culture ($\geq 85\%$) and less consistent when it comes to ethical values, political interest and corruption ($\leq 76\%$), see Appendix A. As all models exhibit similar behavior we focus analysis on mT5-small.

5.2 Cultural value shifts

In Section 5.2.1, we study how the interplay between fine-tuning language, test language and data source will affect the *amount* of value shifts. In Section 5.2.2, we instead test how these factors more generally affect the cultural profiles across test languages by studying in which direction the models' bias changes. Finally, in Section 5.2.3, we test how much cross-lingual sharing during multilingual fine-tuning will further impact these results.

5.2.1 How big is the role of FT language and domain source on cultural value shifts?

In Figure 2, we report the percentage of predictions that remain unchanged after fine-tuning on 2K sentences from news articles, Flores-101, PBC and Tanzil. We find that the amount of changes for Flores-101 are within the same ranges as for the news sources (7-35% shifts). In particular, we find that articles from Reuters (neutral bias) tend to result in the most value shifts. While this is somewhat surprising, it is line with our findings for Flores-101 that show shifts to similar extents. However, as results across these data sources are less distinct in general, we focus most of our further analysis on PBC and Tanzil instead. As expected, we find that PBC and Tanzil have a slightly larger impact on the cultural values encoded (9-43% shifts) than news articles and Flores-101 data. In particular, for PBC, fine-tuning in Korean and Russian have a bigger effect across test languages (e.g. for Greek and German). Similarly, when using Tanzil, next to Korean and Russian, Hindi has a larger effect as well. Yet, similar to our pretrained LMs, we find that the effect of fine-tuning language and source varies across test languages. For instance, for Farsi, regardless of the fine-tuning domain or source, many more values change than for the other languages. This shows that the effect of domain and language bias on the amount of value shifts is heavily dependent on the language that we study shifts for, making it difficult to draw general conclusions on which one has the largest impact overall. We, however, speculate that cultural information is separately encoded for each language in the model, and that the confidence with which these values are encoded varies depending on the test language.[2] Thus, based on the starting point, all fine-tuning setups will be able to affect the test languages to similar extents.

Are certain cultural values more prone to shift? When studying the consistency with which values shift, we find that for each test language the values for the same questions tend to be affected, regardless of the fine-tuning language. Specifically, on average only 14% of the value shifts are unique to only one fine-tuning language and can thus be attributed to language bias. Yet, the values that shift are not consistent across test languages. This again shows that the pattern with which values shift, heavily vary based on the language used for probing. However, these results also indicate that, not only are certain languages more prone to change cultural perspective, there are per language also a specific set of values that are more prone to shift.

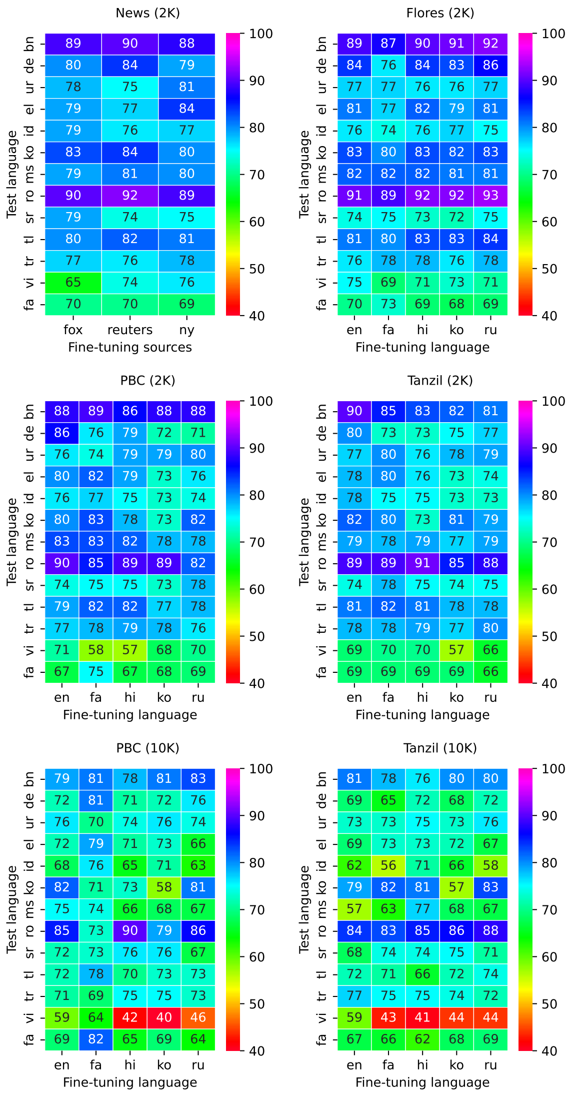


Figure 2: The percentage of predictions that remain unchanged after fine-tuning mT5-small.

Are certain WVS categories more prone to domain bias? In Section 5.1, we found that the pretrained models of different sizes agree more on certain WVS categories. Thus, we test whether the impact of the domain source will be more visible when studying results per category. In Figure 3, we report the Pearson correlation between the percentages of unchanged values per WVS category for each data source pair averaged across fine-tuning languages. We find that overall Tanzil and Bible

[2] If cultural values were jointly encoded across languages, we would expect cultural profiles to behave in a more homogeneous way given the same fine-tuning set up.

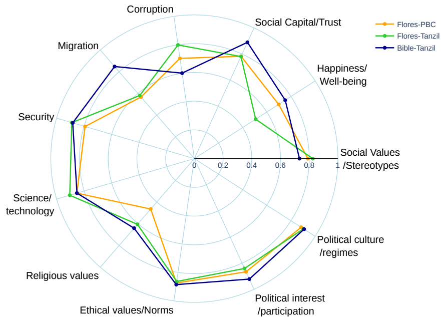


Figure 3: Pearson correlation between the average percentage of unchanged values across fine-tuning languages for each data source pair per WVS category.

tend to score higher compared to Flores-101. Yet, the lowest correlations between data sources are reported for religious values. This suggests that the different religious biases of PBC and Tanzil do in fact have a different effect on the value shifts.

Do we just need more fine-tuning examples? A natural follow up question is whether language or domain bias becomes more prevalent when using a larger corpus during fine-tuning. We find that increasing our training size from 2K to 10K samples does tend to further increase the amount of value shifts (yet, it can also decrease, e.g. for German when fine-tuning in Farsi or Russian). Importantly, however, this does not considerably change the patterns between PBC and Tanzil. Moreover, note that a substantial amount of values shift after fine-tuning on 2K samples, which shows that the fine-tuning procedure has made an impact.

5.2.2 Does the direction of cultural change differ depending on the FT setup?

As we saw that the amount of changes are similar across fine-tuning languages and sources, we now instead study the effect that each fine-tuning setup has on the cultural profiles. In Figure 4, we plot in which direction the cultural profiles of the pretrained model changed depending on the fine-tuning language used. In accordance with our previous results, we find that the direction of change is mostly dependent on the language used for testing, as most fine-tuning languages point in similar directions. However, we do see some differences when comparing results across datasets. For instance, we see that for Indonesian and Korean, the fine-tuning languages seem to steer the cultural bias into different directions depending on the dataset. Moreover, for PBC we find that across many test languages, fine-tuning in Farsi steers the model in a different direction compared to the other languages. Thus, while the amount of value shifts are only weakly affected by fine-tuning language and source, these results indicate that they can have a strong effect on the overall cultural bias of the model.

	bn	de	el	fa	id	ko	ms	ro	sr	tl	tr	ur	vi
mult	76	74	74	72	64	67	72	74	65	69	71	41	78
mono-avg	81	75	74	72	69	71	69	82	73	74	73	48	70

Table 2: Percentage of unchanged values after multilingual and monolingual fine-tuning.

5.2.3 Does multilingual FT affect cultural values differently?

In the previous sections, we studied the effect of monolingual fine-tuning. However, multilingual models are jointly trained on multiple languages, which further complicates which values the model should pick up on. Thus, we now test to what extent cross-language influence during multilingual fine-tuning affects the cultural bias of the models differently compared to monolingual fine-tuning. In Table 2, we report the amount of unchanged values after multilingual and monolingual fine-tuning. We find that the effect of multilingual fine-tuning is for many languages approximately similar to the average scores obtained across fine-tuning languages in a monolingual set-up (this pattern holds for models of varying sizes, see Appendix C). In addition, in Figure 4 we see that the direction in which the cultural profiles change, does not deviate much from monolingual fine-tuning in most cases. We suspect that the results for multilingual fine-tuning are similar because the fine-tuning languages in any case tend to behave similarly. Thus, when using them interchangeably it has a limited further effect on the predictions.

6 Correlation with human survey results

In Section 5.1, we studied which cultural values were encoded in the pretrained LM, and in Section 5.2, how much and in which direction these could change after fine-tuning. We now test whether the changes we observed led the model to be steered into a direction that is better aligned to real human values. Thus, we compute how much the Spearman correlation between the ground truth profiles and the pretrained LM changed after fine-tuning. To select culturally diverse countries to test alignment to, we first compute cosine similarities between the ground truth profiles of our 13 test

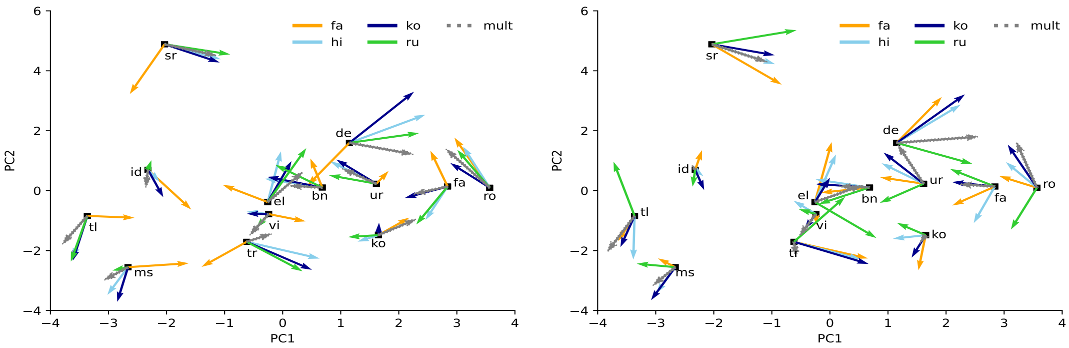


Figure 4: Starting from the cultural profiles extracted from pretrained mT5-small, the image depicts in which direction each test language changes depending on the language selected for fine-tuning on PBC (left) and Tanzil (right). The cultural profiles are projected down to 2-dimensions using PCA (Bro and Smilde, 2014).

	bn	de	ur	el	id	ko	ms	ro	sr	tl	tr	vi	fa
							PBC						
DE	+.02	+.11	+.03	+.04	+.09	-.03	+.12	+.01	-.05	+.12	+.04	+.15	-.03
PK	-.13	+.02	-.11	+.16	+.14	+.13	+.18	-.08	+.01	+.23	+.15	-.02	-.14
SR	-.06	+.26	-.06	+.08	+.04	-.01	+.01	0	-.08	+.16	+.09	0	-.07
VI	-.08	-.18	-.11	0	+.01	-.01	-.03	-.09	+.01	+.12	0	-.14	+.04
							Tanzil						
DE	-.02	+.18	+.06	+.02	+.10	-.04	+.12	+.08	-.06	+.14	-.02	+.19	-.02
PK	-.16	+.02	-.08	+.16	+.12	+.13	+.11	-.08	0	+.26	+.18	-.01	-.14
SR	-.08	+.03	-.06	+.09	+.03	-.04	-.01	0	-.06	+.15	+.05	+.04	-.06
VI	-.09	-.24	-.07	+.01	-.08	+.05	-.08	-.11	+.05	+.11	+.04	-.15	+.05

Table 3: Change in alignment to the ground truth profiles for each country (DE, PK, SR, VI), measured by the difference in Spearman correlation. Results are averaged over fine-tuning languages.

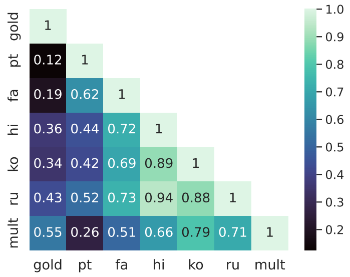


Figure 5: Spearman correlation between the similarity matrices of the cultural profiles computed from the ground truth data, and pretrained and fine-tuned models.

languages, and found that the profiles from Germany and Pakistan (0.84 similarity) and Vietnam and Serbia (0.88 similarity) deviated the most.

How does the alignment between test languages and human values change? In Table 3, we report the change in correlations averaged over fine-tuning languages. In Figure 4 we saw that, depending on the test language, fine-tuning changed the cultural information in different directions. We now see that this mostly leads to a better alignment to human data, regardless of domain source. For instance, fine-tuning on Tanzil and PBC on average increases correlation with all countries' data for Tagalog. This suggests that the model is overall pushed closer to human values. Moreover, test languages whose profiles pointed in different directions across datasets (e.g., Korean, Indonesian and Serbian), are now also affected differently in their alignment to human values. Yet, when looking at the absolute values for PBC and Tanzil, the correlations remain low across the board, showing that the model is still poorly aligned to human values.

How does fine-tuning affect the cultural similarity between test languages within a model? Given that our models are poorly aligned to human data, we test whether at least the cultural similarities between different test languages correlate with those between real countries. For instance, do we find that the cultural profiles from Romanian and Serbian are more similar than those from Serbian and Urdu? To test this, we, for each model, compute a dissimilarity matrix between the cultural profiles of all language pairs using cosine similarity, and then use Spearman correlation to test how similar these matrices are across models (Abnar et al., 2019). In Figure 5, we find that while cultural relationships between languages in the pretrained LM are weakly correlated with human data, the alignment mostly increases after fine-tuning on PBC, (except for Farsi). Interestingly, multilingual fine-tuning results in the highest correlation with human data. The same result was found for Tanzil, see Appendix D. When looking at the dissimilarity matrices for these models, we also find that the cultural profiles are more distinct, resulting in less similarity between language pairs. We suspect that, as a result of seeing multiple languages during training, various language-specific biases can be preserved and transmitted. In contrast, after monolingual fine-tuning, all languages are biased in one direction, resulting in very similar cultural profiles that do not preserve cross-cultural differences.

7 Tracing cultural value shifts

We found that fine-tuning languages have similar effects across test languages. Thus, as a complementary study, we test which training examples, and the languages they come from, influenced value shifts the most. We use TRAK, a TDA method proposed by Park et al. (2023). We follow Park et al. (2023) in treating the MLM objective as a multi-class classification problem, i.e., framing it as a sequence of v-way classification problems over masked tokens, where v is the vocabulary size. We use the TRAK library[3] for our implementation and project gradients down to 4096 dimensions, all other parameters are kept at default. See Park et al. (2023) for a detailed explanation. Per value shift we analyze the top 100 most influential training examples.

Are value shifts influenced by the same training examples across fine-tuning and test languages? In Section 5.1, we saw that value shifts were mostly not unique to one fine-tuning language. Thus, we test whether these shifts were actually influenced by the same training examples across fine-tuning languages. Interestingly, we find that from the most influential examples in each fine-tuning language that instigate the same value shift, only <5% are parallel sentences. Yet, when looking at the values that shift across test languages given the same fine-tuning language, we observe more consistency. For PBC, we find that across all language pairs, when fine-tuning in Farsi and Hindi, up to 20% of training examples are consistently relied upon across test languages, and for Korean and Russian 49 and 62% resp. These results suggest that the semantic content of fine-tuning data might not be the main reason behind the shifts. Instead, the model tends to rely more on the same training examples within a fine-tuning language irrespective of test language.

Which languages instigate the value shifts during multilingual fine-tuning? We use the approach from Choenni et al. (2023a) to quantify cross-language influence by the average percentage of training examples that each fine-tuning language contributes to the most influential examples for each test language. In Figure 6, we see that for PBC, Russian and Farsi have on average the largest influence across test languages. Interestingly, for Tanzil, we instead see that Russian and Korean contribute the most. Given that different trends across datasets could also be an artifact of randomness during fine-tuning, we repeat the experiment for PBC on a model fine-tuned with a different random seed, but confirm that the trends hold. While the large influence of Russian could be explained by the fact that it has the second largest dataset for pretraining, the results indicate that the influence of languages on value shifts during multilingual fine-tuning is dependent on the content of the fine-tuning data.

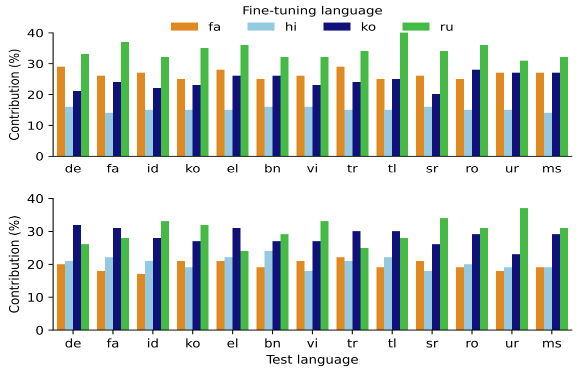


Figure 6: The average percentage of training samples from each fine-tuning language that contributed to the top 100 training samples for a test language after multilingual fine-tuning on PBC (top) and Tanzil (bottom).

8 Discussion and conclusion

We studied to what extent fine-tuning languages and domain sources exert influence on cultural values encoded for a set of test languages in MLMs. In particular, we tested how different fine-tuning setups can change the overall cultural biases across test languages differently, and in which cases this leads the model to be better aligned to human values. We found that fine-tuning language and domain source play a minor, but visible, role in the amount of value shifts compared to size of the fine-tuning dataset. Moreover, results vary considerably across test languages. Still, different fine-tuning languages can cause cultural profiles of test languages to be steered into different directions, which leads to varying effects on the models' alignment to human values. In addition, we find that multilingual fine-tuning better preserves the human cultural similarities between test languages within a LM.

[3] https://github.com/MadryLab/trak

Finally, our TDA analysis shows that while different fine-tuning languages can lead to the same value shifts, the training examples that are relied upon vary. This suggests that the semantic content of fine-tuning data might no be the main reason for the shifts. Instead, the model tends to rely on the same training examples within a fine-tuning language, and these examples have different effects on the manifestation of cultural values across test languages. Hence, future work on value alignment likely requires a different adaptation approach for each test language. While multilingual NLP has made big strides in the past years, the field of cross-cultural NLP is still in its infancy as many questions remain to be explored. We hope that our insights will inform future work on value alignment to enable more culturally-aware language technology.

Limitations

While language and culture are closely connected (Kramsch, 2014; Hovy and Yang, 2021), we can not use these notions interchangeably (Hershcovich et al., 2022b). For instance, even within a language many subcultures typically exist, and the idea that for instance "English" carries a single set of values has been discarded (Paul and Girju, 2009). At the same time, multiple languages can also carry a relatively homogeneous culture (Sahlgren et al., 2021). While the languages were selected based on the criteria that its speakers can be primarily localized to a specific geographical region (and thus likely maintain their own cultural profile), we can not guarantee that all online texts in that language transmit the same cultural values.

Moreover, we were restricted in the choice of domain source and fine-tuning language combination due to a lack of available datasets that contain a sufficient amount of multi-parallel data for fine-tuning. While we could, for instance, use many languages from the Flores-101 dataset, each language only contains approximately 2K multi-parallel sentences. While PBC and Tanzil contain different religious biases, it could also be argued that these data sources are in fact not substantially dissimilar.

Finally, while we use data from one of the most popular cross-cultural value questionnaires from social science, i.e. WVS, it also has its shortcomings. In particular, similar to how languages do not contain a single culture, it is also questionable to map an entire country to a single set of cultural values. This is particularly true for countries with many immigrants of different cultural backgrounds. As such, there are also different subcultures within a country, making it not obvious that we should simply map a MLM to a countries' cultural values based on the WVS data. This also further complicates how we should interpret an alignment between language and country as it can easily be mismatched. Thus, in future work, researchers from various disciplines should investigate and discuss what an ideal cultural alignment for a MLM should look like in practice.

Ethical considerations

All data sources used in this study are publicly available. While we acknowledge that automatically analyzing religious texts can be a sensitive topic, we do not draw any conclusions based on the content of those data sources in this work nor do we provide examples from the texts directly. Moreover, while we test for cultural alignment to human data in this study, we recognize that languages can not simply be mapped to single countries and therefore it is not always straightforward to decide which human values the model should align to in practice. As such, we leave this question in the middle, and rather just explore to what extent we can influence the cultural profiles of MLMs.

Acknowledgements

Rochelle Choenni is supported by a Google PhD Fellowship. The work of Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and the Federal States.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Jelle Zuidema. 2019. Blackbox Meets Blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains. In *Proceedings of the ACL-Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 191–203.

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards Tracing Knowledge in Language Models Back to the Training Data. *arXiv preprint arXiv:2205.11482*.

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.

Rasmus Bro and Age K Smilde. 2014. Principal Component Analysis. *Analytical methods*, 6(9):2812–2831.

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.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023a. How do languages influence each other? studying cross-lingual data sharing during lm finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13244–13257.

Rochelle Choenni, Ekaterina Shutova, and Dan Garrette. 2023b. Examining Modularity in Multilingual LMs via Language-Specialized Subnetworks. *arXiv preprint arXiv:2311.08273*.

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1491.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2021. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10333–10350.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bi Puranen, et al. 2022. World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, 12(10):8.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Constantin A Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. Speaking multiple languages affects the moral bias of language models. *arXiv preprint arXiv:2211.07733*.

Xiaochuang Han and Yulia Tsvetkov. 2022. ORCA: Interpreting Prompted Language Models via Locating Supporting Data Evidence in the Ocean of Pretraining Data. *arXiv preprint arXiv:2205.12600*.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual Language Models are not Multicultural: A Case Study in Emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214.

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. 2022a. 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.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022b. Challenges and Strategies in Cross-Cultural NLP. In *60th Annual Meeting of the Association-for-Computational-Linguistics (ACL), MAY 22-27, 2022, Dublin, IRELAND*, pages 6997–7013. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender Bias in Masked Language Models for Multiple Languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750.

Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves

Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Claire Kramsch. 2014. Language and culture. *AILA review*, 27(1):30–55.

Tsz Kin Lam, Eva Hasler, and Felix Hieber. 2022. Analyzing the Use of Influence Functions for Instance-Specific Data Filtering in Neural Machine Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 295–309.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.

Yuxian Meng, Chun Fan, Zijun Sun, Eduard Hovy, Fei Wu, and Jiwei Li. 2020. Pair the Dots: Jointly Examining Training History and Test Stimuli for Model Interpretability. *arXiv preprint arXiv:2010.06943*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Tarek Naous, Michael J Ryan, and Wei Xu. 2023. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *arXiv preprint arXiv:2305.14456*.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*.

Michael Paul and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1408–1417.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Magnus Sahlgren, Fredrik Carlsson, Fredrik Olsson, and Love Börjeson. 2021. It's basically the same language anyway: the case for a Nordic language model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 367–372.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2023. Quantifying gender bias towards politicians in cross-lingual language models. *Plos one*, 18(11):e0277640.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Alexandra Sasha Luccioni10, Maraim Masoud11, Margaret Mitchell10, Dragomir Radev12, et al. 2022. You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings. *Challenges & Perspectives in Creating Large Language Models*, page 26.

Wilson L Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, Ethics, Morals? On the Use of Moral Concepts in NLP Research. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From Instructions to Intrinsic Human Values–A Survey of Alignment Goals for Big Models. *arXiv preprint arXiv:2308.12014*.

A Agreement between pretrained LMs

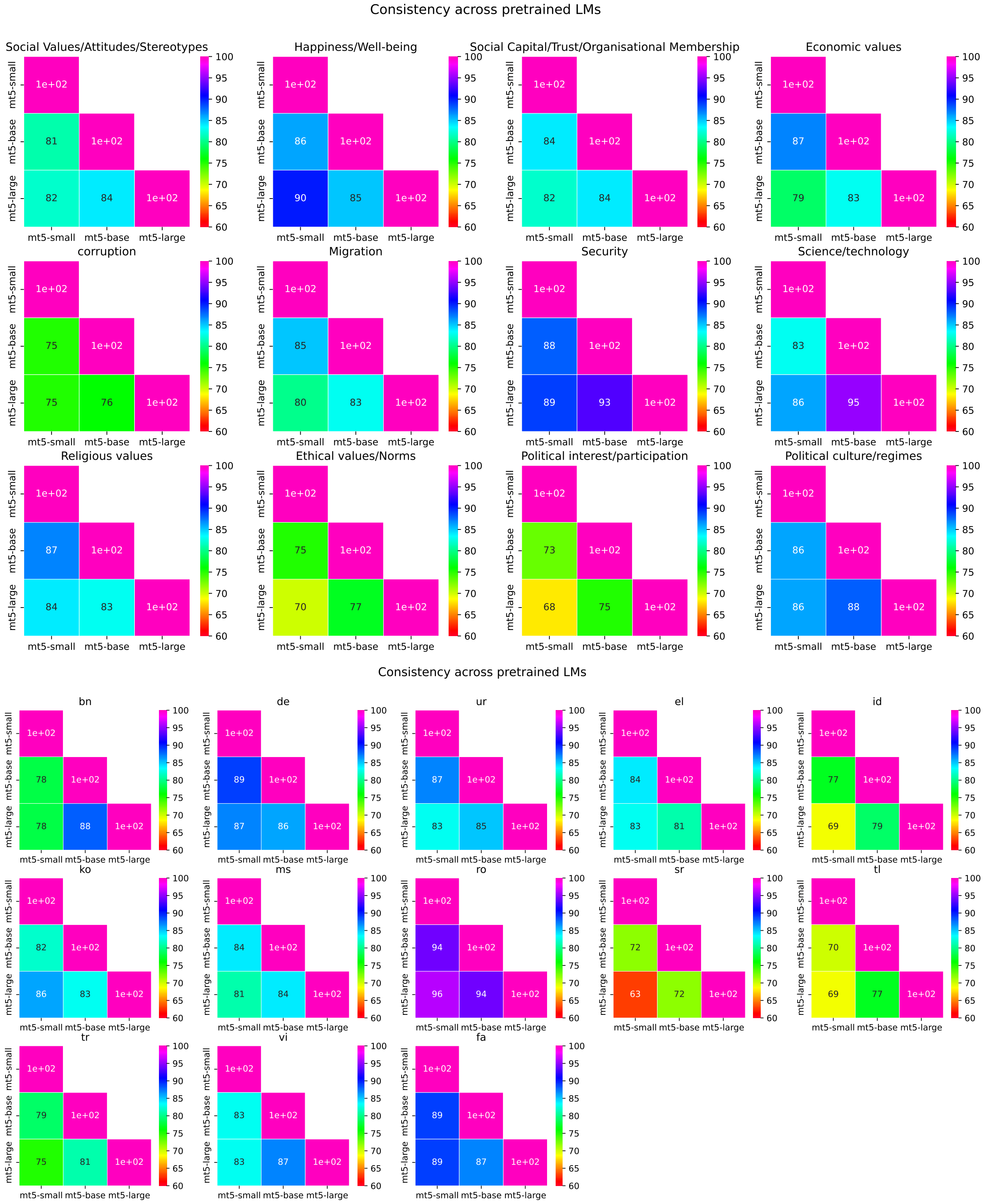


Figure 7: The percentage of survey questions for which pretrained mT5-small models with different number of parameters were in agreement about the answer they outputted. We show the percentage per category averaged over test languages (top) and the percentage per test language averaged over categories (bottom)

B Fine-tuning details

We use a 80/20 train/development split, a learning rate of 5e-5, the AdamW optimizer and a batch size of 8, and train for 5 epochs. We query the models through the Huggingface Library and use its Trainer class with default hyperparameters for fine-tuning (Wolf et al., 2019). Moereover, all fine-tuning and tracing experiments are ran on a NVIDIA A100-SXM4 GPU with 40GB memory.

C Percentage of unchanged value predictions after fine-tuning

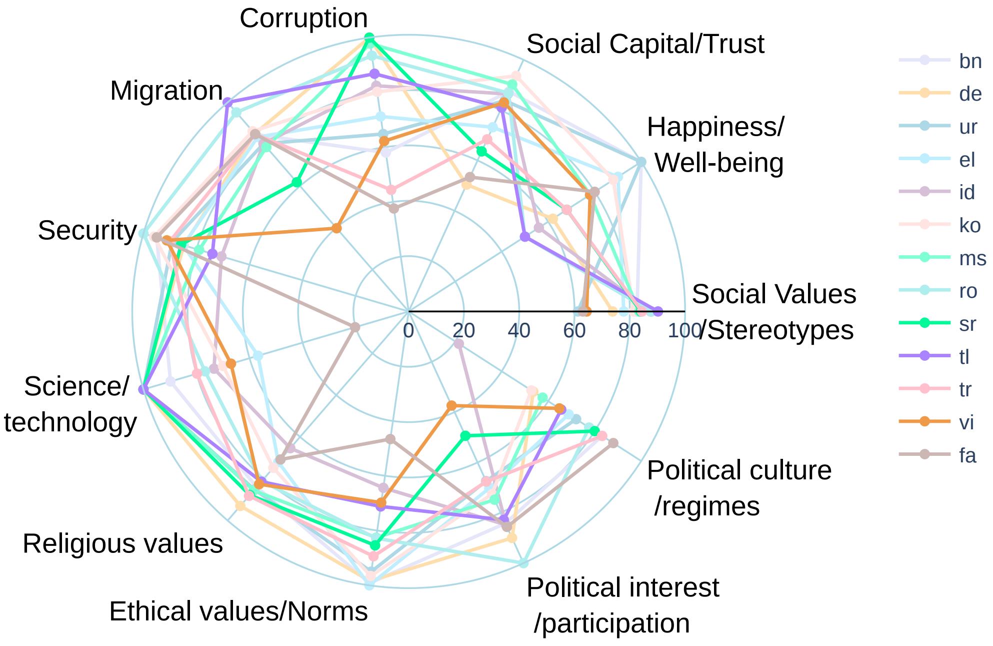


Figure 8: The percentage of unchanged values per test language for each WVS category after fine-tuning on PBC. Results are averaged across fine-tuning languages.

Robustness analysis We separately fine-tune the mT5-small in each language on PBC with 3 random seeds, and show results of two seeds in Figure 9. We find that both across different random seeds for the same model and across mT5 of different model sizes, the amount of predictions that change after fine-tuning compared to the pretrained LM are relatively similar. However, we do see that for mT5-large, language-wise patterns become more distinct. For instance, across all fine-tuning languages, we see that the predictions for Bengali and Urdu remain more robust compared to the smaller models. For Turkish and Indonesian we see an opposite effect where instead across all fine-tuning languages the predicted values tend to change more. Similarly, we compared performance to fine-tuning using only 1 training epoch, while this slightly reduces the amount of value shifts, the overall patterns did not change considerably.

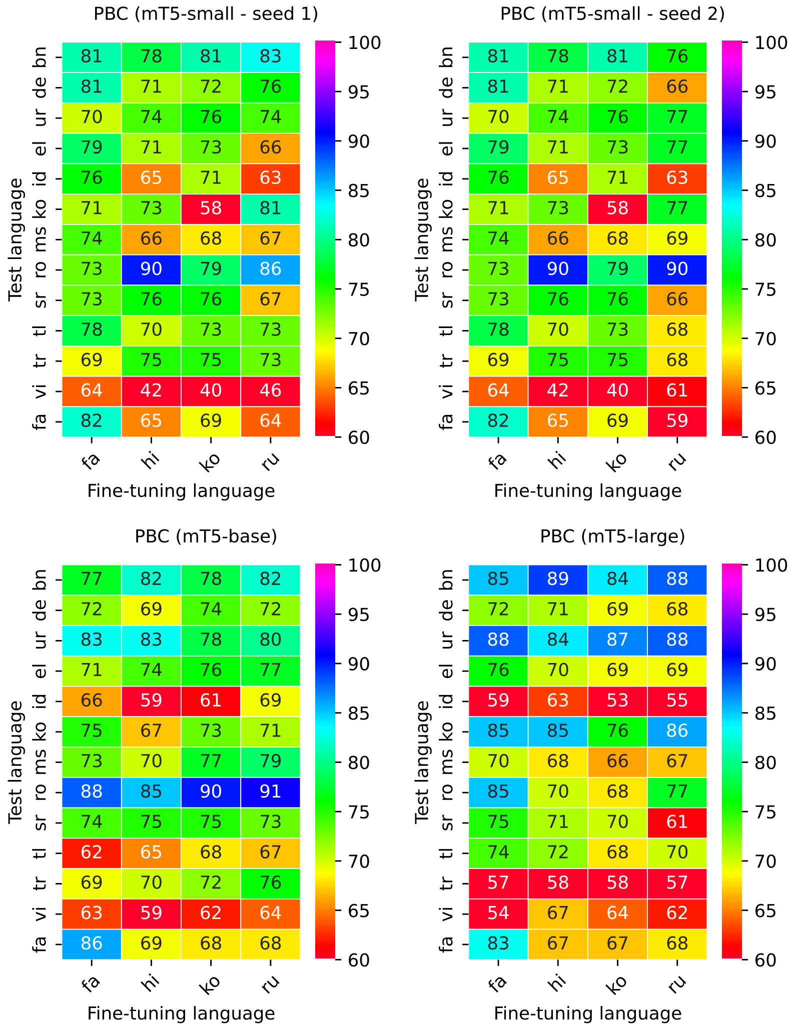


Figure 9: The percentage of of unchanged values after fine-tuning mT5-small on 10K sentences from PBC. We see the effect of using 2 different random seeds during fine-tuning, and the effect of using different model sizes i.e., mT5-base and mT5-large.

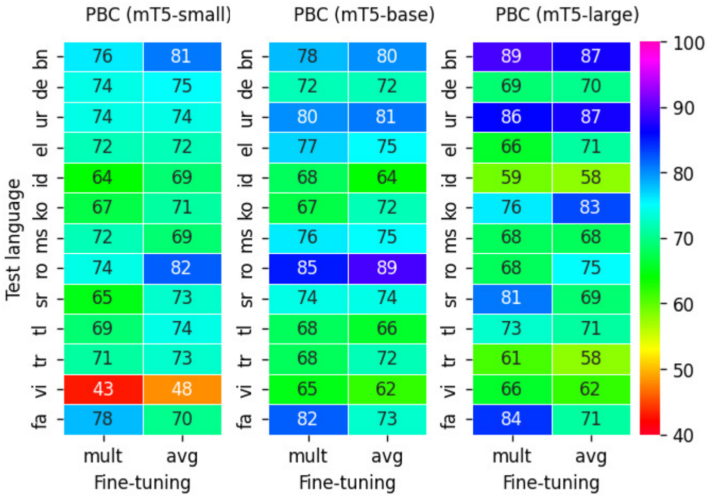


Figure 10: The percentage of survey questions for which the value prediction did not change after multilingual fine-tuning and on average when monolingual fine-tuning using the same languages.

D Changes to cultural profiles after fine-tuning

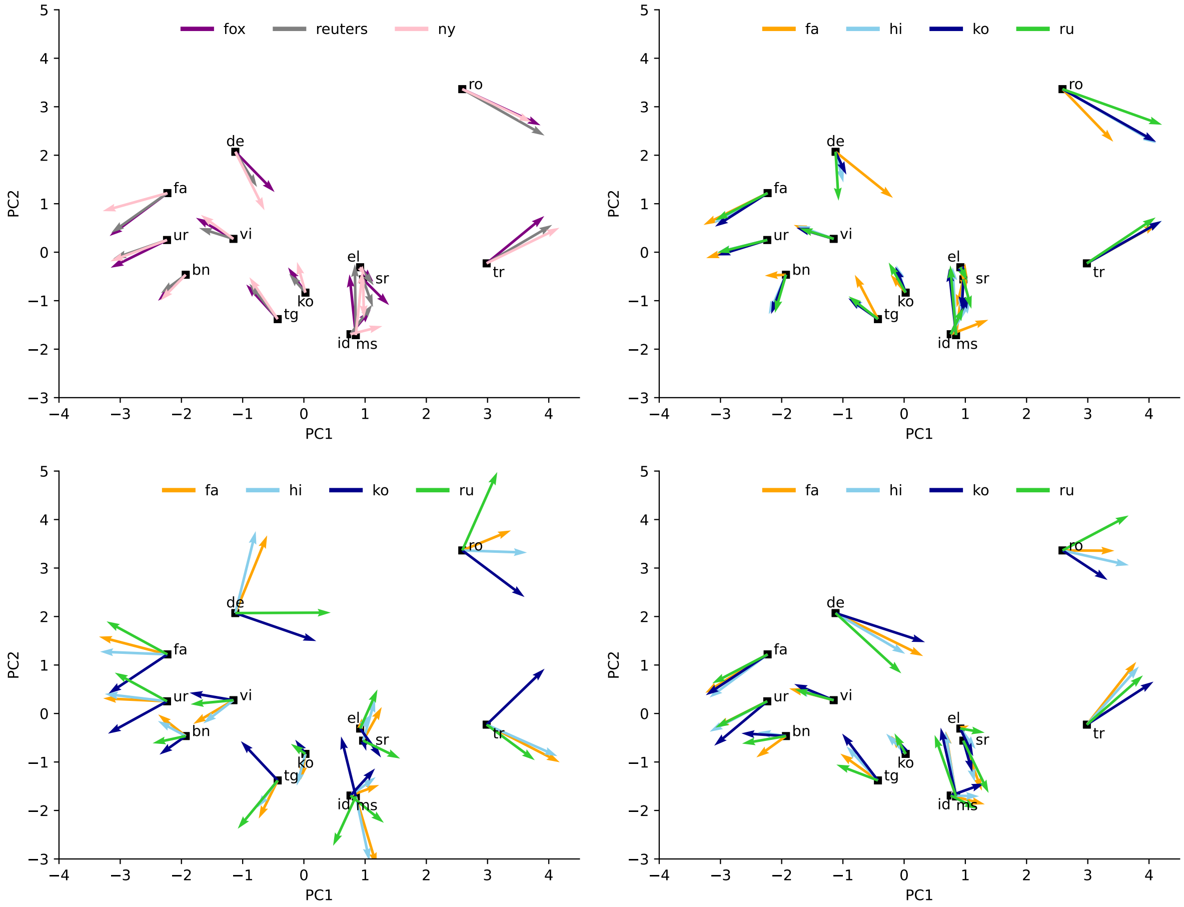
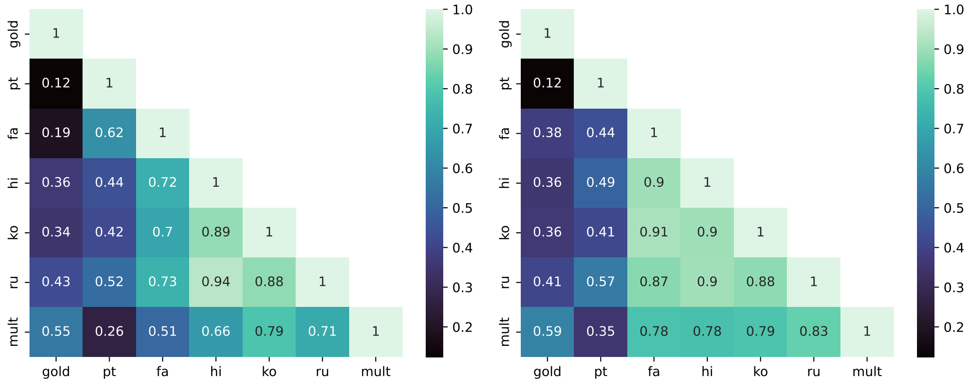


Figure 11: Starting from the cultural profiles extracted from pretrained mT5-small, the image depicts into which direction each test language changes depending on the source selected for fine-tuning on 2K sentences: news articles (top left), Flores (top right), PBC (bottom left) and Tanzil (bottom right. The cultural profiles are projected down to 2-dimensions using PCA.



	gold	pt	fa	hi	ko	ru	mult
gold	1						
pt	0.12	1					
fa	0.19	0.62	1				
hi	0.36	0.44	0.72	1			
ko	0.34	0.42	0.7	0.89	1		
ru	0.43	0.52	0.73	0.94	0.88	1	
mult	0.55	0.26	0.51	0.66	0.79	0.71	1

	gold	pt	fa	hi	ko	ru	mult
gold	1						
pt	0.12	1					
fa	0.38	0.44	1				
hi	0.36	0.49	0.9	1			
ko	0.36	0.41	0.91	0.9	1		
ru	0.41	0.57	0.87	0.9	0.88	1	
mult	0.59	0.35	0.78	0.78	0.79	0.83	1

Figure 12: Spearman correlation between the similarity matrices of the cultural profiles computed from the ground truth data, and pretrained and the models fine-tuned on PBC (left) and Tanzil (right).

E Ground truth cultural profiles

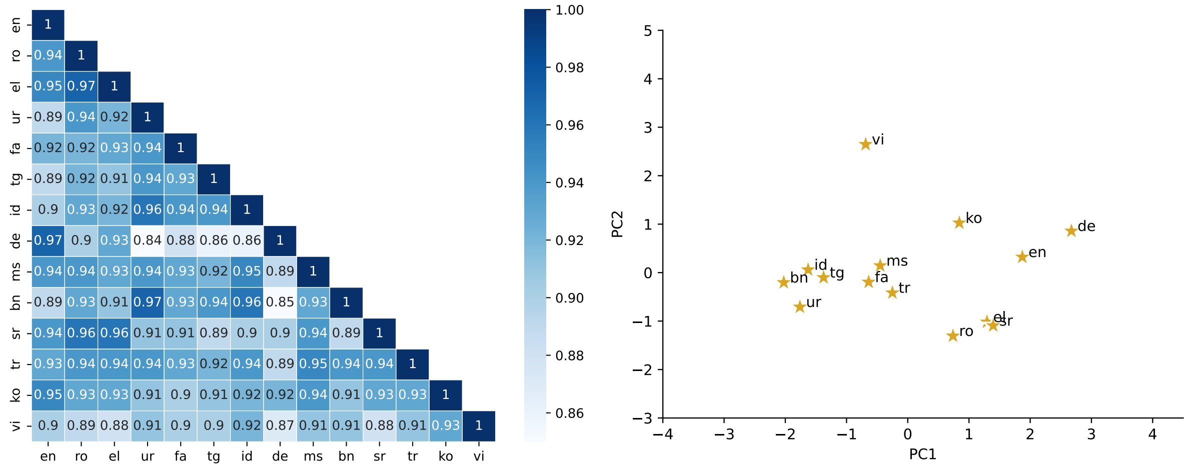


Figure 13: Left: The similarity between the cultural profiles of different countries according to the WVS survey results. Right: the ground truth profiles from each country projected down to 2 dimensions using PCA.

F TRAK analysis

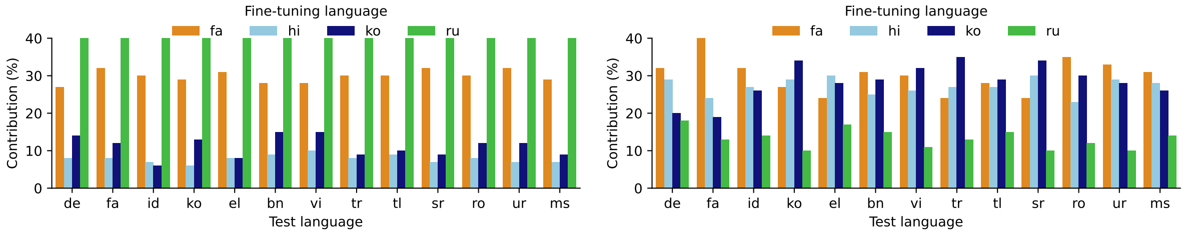


Figure 14: The average percentage of training samples from each fine-tuning language that contributed to the top 100 *contradicting* training samples for a test language after multilingual fine-tuning on PBC (left) and Tanzil (right).