Soft Prompts for Adapting LLMs to Cultural Commonsense Knowledge

Gabrielle Le Bellier[1], Marine Carpuat[2], Benoît Sagot[1], Chloé Clavel[1]
[1]Inria, Paris, France, [2]University of Maryland, College Park, United States
{gabrielle.le-bellier, benoit.sagot, chloe.clavel}@inria.fr, marine@umd.edu

Abstract

Large Language Models (LLMs) show unbalanced knowledge of cultures across the globe, favoring high-resource cultures over low-resource ones. A possible way to tackle this issue is to fine-tune LLMs on culturally specific data. However, fine-tuning recent LLMs requires high computational resources as well as memory storage, which triggered the development of parameter-efficient fine-tuning (PEFT) approaches, the most widespread being LoRA. In this article, we investigate the use of another class of PEFT approaches, namely soft prompt methods (prompt-tuning and prefix-tuning), to improve LLMs' cultural knowledge across diverse cultures. We focus on cultural alignment on Multiple-Choice Questions of cultural commonsense knowledge. On this task with limited fine-tuning data, we show that soft-prompt-based methods outperform LoRA in comparable settings. Moreover, the trained soft prompts are interpretable and capture similarities between cultures.[1]

1 Introduction

Large Language Models (LLMs) exhibit cultural biases and struggle to understand and apply cultural norms, values, and commonsense when asked to perform culturally sensitive tasks. Recent research in NLP focuses on the evaluation of cultural awareness of models, defined as the ability to understand the cultural aspects of a given context and apply relevant behavior to perform a task in that context (Guo et al., 2025; Pawar et al., 2025a). LLMs are shown to lack cultural awareness, especially towards low-resource cultures, favoring the WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations (Henrich et al., 2010).

Moreover, there is a common effort towards cultural alignment. Cultural alignment in models refers to their capacity to reflect the diverse aspects of life (*e.g.*, views, beliefs, knowledge, and habits) characteristic of a given cultural group (AlKhamissi et al., 2024; Orlowski et al., 2025). Various approaches have been proposed to develop culturally aligned models, including fine-tuning on culturally relevant data. To represent culture, existing datasets typically rely on concrete proxies, such as norms, values, traditions, food, and commonsense knowledge. However, these cultural aspects cannot be treated uniformly by models. Norms and values are often subject to significant variation across individuals within the same cultural group, whereas commonsense cultural knowledge is generally assumed to be more consistently shared, regardless of factors such as age, gender, or socioeconomic status (Nguyen et al., 2024).

However, as the size of models keeps growing, fine-tuning large models becomes more computationally expensive and requires large memory space. Therefore, some solutions have arisen in recent years to tackle this issue. Parameter-Efficient Fine-Tuning (PEFT) is a collection of methods that enable fine-tuning only a subset of parameters while keeping the majority of the model frozen, which makes these approaches computationally and memory efficient. The reparameterization method LoRA (Hu et al., 2021) is the most widely adopted PEFT approach and has been shown to achieve performance close to full fine-tuning while requiring significantly lower computational and memory resources. Other PEFT methods emerged, such as soft prompts methods which add a sequence of trainable parameters in the input, whether only at the initial layer in the case of prompt-tuning (Lester et al., 2021), whether at each layer for prefix-tuning (Li and Liang, 2021). Once trained, the soft prompts capture the way the models learn cultural aspects, and can thus be used for interpretability purposes.

In this paper, we investigate cultural adaptation through soft prompt methods. More precisely,

[1]Our codebase is publicly available here.

Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026), pages 76–100

for each considered culture, we aim to derive a lightweight model specific to the cultural commonsense knowledge of that culture. We compare prompt-tuning and prefix-tuning with the widely used LoRA method. Moreover, by studying the trained soft prompts, we expect to interpret the relationships and similarities between cultures.

Our main contributions are threefold. Firstly, we find that soft prompt methods significantly improve cultural alignment despite involving very few parameters, and outperform LoRA when used in comparable settings. Secondly, we show that fine-tuned models also outperform the baseline on cultures other than the one used for fine-tuning, particularly on high-resource ones. Finally, we show that the relative positioning of the resulting soft prompts in the embedding space can serve as a basis for quantifying cultural similarity as learned by the model.

2 Related Work

2.1 Cultural awareness in NLP

2.1.1 Culture and cultural awareness

In anthropology, culture is defined in multiple ways and remains a subject of ongoing debate among anthropologists. A widely cited definition is that of E.B. Tylor, who characterizes it as "that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society." (Tylor, 1871). Indeed, culture is learned and perpetuated within a cultural group, while simultaneously being continuously questioned and evolving over time. A culture is embodied in norms, values, symbols, and mental maps, but some anthropologists argue that culture is not merely an accumulation of its parts. In this sense, Geertz (1973) asserts that culture is, rather, "webs of significance."

In NLP, culture has become an increasingly studied research topic (Liu et al., 2025a). Researchers are examining the cultural awareness of models, defined as a model's ability to understand a given cultural context and take it into account when performing a specific task (Pawar et al., 2025a). In practical terms, this amounts to applying cultural norms to social situations or correctly answering culture-dependent questions. For example, in the context of greetings, a model aligned with French culture would refer to cheek kissing, whereas a model aligned with Japanese culture would refer to bowing. A model's alignment with a cultural group is measured by the concordance between its responses and those provided by individuals from the corresponding cultural group (AlKhamissi et al., 2024; Durmus et al., 2024), for example in value surveys (Pew Research Center, 2026; Haerpfer et al., 2022).

To embed the abstract notion of culture in NLP, existing work relies on concrete dimensions that can be encoded in datasets for evaluating and improving cultural alignment. These dimensions include, among others, norms and values (Ramezani and Xu, 2023), food habits (Hu et al., 2024), word associations (Liu et al., 2025c; Dai et al., 2025), or commonsense knowledge (Nguyen et al., 2023, 2024; Yin et al., 2022). However, these dimensions can not be handled similarly by models. In particular, norms and values are often subjective, context-dependent, and influenced by individual characteristics such as gender, age, or socioeconomic status, making them difficult to generalize across an entire cultural group. In contrast, cultural commonsense knowledge typically reflects more widely shared assumptions within a given cultural group and is therefore more prone to generalization at the group level.

2.1.2 Cultural commonsense knowledge

Multiple datasets gather cultural commonsense knowledge. CANDLE (Nguyen et al., 2023) is a dataset of assertions extracted from the Internet (via the C4 Web Crawl) and written in English. Nguyen et al. (2024) expanded the number of sentences by prompting GPT-3.5 using previous datasets (CANDLE and CONCEPTNET (Liu and Singh, 2004)) to create MANGO. GEOMLAMA (Yin et al., 2022) contains sentences with masked words, as well as multiple-choice questions and answers, in five different languages to assess the general cultural knowledge of models from five countries. The collection of cultural data thus faces a major problem: there are few human annotators, particularly for underrepresented cultures. Datasets are therefore limited in size, which can compromise cultural fine-tuning or robust evaluation, or are generated using LLMs, which can introduce culturally stereotyped content in the absence of human supervision or validation.

To address the lack of massive cultural knowledge data annotated by humans from the cultures they are annotating, BLEND (Myung et al., 2024) compiles 52.2K question-answer pairs at the granularity level of a country or region. The questions

are provided in the corresponding languages, along with their English translations.

2.1.3 Cultural alignment

Beyond the evaluation of models across multiple cultural aspects, studies have focused on aligning models with specific cultures, particularly underrepresented ones (Putri et al., 2024; Etxaniz et al., 2024). In the current state of the art, models can be culturally aligned using several techniques. At first glance, anthropological prompting (*e.g.*, gender, country, occupation, social class) appears to enhance model alignment with specific human groups (AlKhamissi et al., 2024; Tao et al., 2024). However, this approach tends to produce stereotypical outputs for certain cultures, thereby reflecting and potentially amplifying the cultural biases embedded in the models (Pawar et al., 2025b; Durmus et al., 2024). To counter this drawback, models require external cultural information, which can be incorporated in various ways.

RAG (Retrieval Augmented Generation) can be seen as a way to tackle the lack of cultural knowledge of models. Nguyen et al. (2023) and Lertvittayakumjorn et al. (2025) show that retrieving cultural facts improves cultural alignment. However, these methods are slow at decoding time and do not result in culture-specific models.

Therefore, most efforts in cultural alignment have been put towards fine-tuning LLMs on a target culture embodied by a culture-relevant dataset. Yao et al. (2025) employ full fine-tuning on their conversational cultural data to produce culturally aligned conversational models on five different cultures. Several studies rely on fine-tuning on Supervised Fine-Tuning (SFT) and human preference optimization (Feng et al., 2025; Guo et al., 2025). However, fine-tuning all the weights of an LLM is demanding in computational resources and memory capacities. To enable fine-tuning at lower cost, cultural alignment often relies on LoRA (Hu et al., 2021), a well-known PEFT method (Li et al., 2024; Adilazuarda et al., 2025; Liu et al., 2025b). A more recent approach considers Mixtures of Experts combined by LoRA adapters (Sun et al., 2026). Furthermore, other PEFT methods are sometimes used: Yang et al. (2023) use continuous prompts to control the type of food (Mexican, Asian, etc.) in the generation of restaurant reviews, while Masoud et al. (2024) train continuous prompts with a black-box optimization method for cultural survey alignment.

In our study, we aim to develop models that are computationally efficient during inference, with each model being culture-specific. We therefore choose to use PEFT methods.

2.2 Parameter-Efficient methods

As we have just seen, cultural alignment sometimes requires adjusting the model's parameters on cultural data. However, adjusting all parameters requires substantial computational and memory resources: this is a problem faced by the entire NLP community, beyond cultural considerations alone. To address the high resource requirements of full fine-tuning, recent work has focused on more efficient and less resource-intensive approaches. PEFT (Parameter-efficient Fine-tuning) methods allow results close to those of full model fine-tuning to be achieved at a lower cost (Han et al., 2024; Zhang et al., 2025). These methods are based on fine-tuning a small subset of parameters, while the rest of the model remains frozen.

Among PEFT methods, we focus specifically on methods using continuous prompts (soft prompts), where continuous tokens are trained while the model remains frozen. Intuitively, this amounts to optimizing the prompt that would inject cultural knowledge into the model, without the biases associated with the user's choice of prompt, a choice that can significantly vary the model's performance. Prompt-tuning (Lester et al., 2021) introduce continuous prompts, that is, a sequence of virtual tokens that can be adjusted in the input sequence. Prefix-tuning (Li and Liang, 2021) suggest applying continuous tokens to all layers of the model, by adding trainable parameters to the keys and values of the attention mechanism in each layer. In recent years, methods derived from soft prompts have emerged (Liu et al., 2022, 2024; Wang et al., 2025; Zhang et al., 2023; Qian et al., 2022; Wang and Demberg, 2024). However, continuous prompts are difficult to compare with human-written discrete prompts (hard prompts). The similarities (*e.g.*, cosine similarities) between continuous and discrete prompts are uninterpretable (Lester et al., 2021; Khashabi et al., 2022). However, we can interpret continuous prompts by using the similarities between them. In this way, Vu et al. (2022) explore similarities between different tasks by observing the similarities between the continuous prompts trained on each task. The authors achieve more effective task transfer, where they can leverage a prompt trained on a specific task to adapt it to another closely related task.

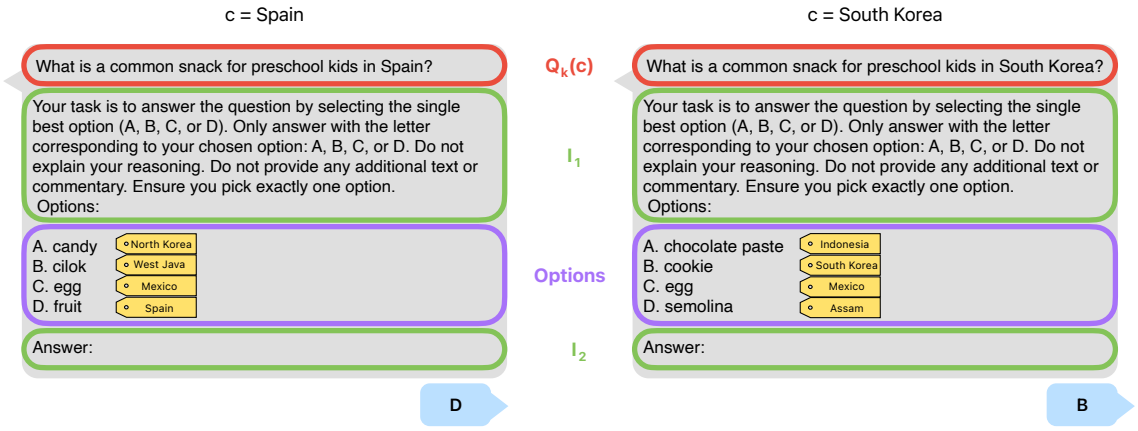


Figure 1: Two examples from the BLEND dataset (MCQ part), taken from the question template "What is a common snack for preschool kids in <country>?". For each example, the four options are randomly drawn from four different cultures (shown in the Figure, but not included in the data), including the correct answer. All BLEND data we use is in English.

In our study, we aim to use the most widely used soft prompt methods to fine-tune models on specific cultures. Unlike hard prompts, which depend on user-defined formulations, soft prompts are randomly initialized and optimized on data representing the target culture, without requiring an explicit characterization of the culture. By sharing a single base model and learning one prompt per culture, memory requirements are significantly reduced, in contrast to storing a full model for each culture. Furthermore, we aim to exploit the trained prompts to investigate how the model encodes the culture. We will compare soft prompt methods with the widely used LoRA approach.

3 Methodology

To address the lack of cultural knowledge of a model, we fine-tune it on culture-specific data. For a given culture, we consider the corresponding dataset of knowledge composed of a culture-dependent number of question-answer pairs. We fine-tune the model $\mathcal{M}$ on this dataset to produce the culture-specific model $\mathcal{M}_c$. We employ soft prompt methods, which require the training of a very small subset of parameters compared to the size of the model. Our approach can be seen as a controllable generation task, where the control attribute is the culture embedded in the soft prompt.

3.1 Data

3.1.1 Multiple-Choice Questions

BLEND (Myung et al., 2024) is a dataset with granularity at the country level (*e.g.*, Spain, Iran) or region level (*e.g.*, West Java). The question templates were created by annotators targeting facets of their cultures (*e.g.*, "What is a common snack for preschool kids in your country?"). The facets are divided into six topics: *Food*, *Education*, *Holidays/Leisure/Celebrations*, *Work life*, *Sport* and *Family*. Once collected, these question templates were submitted to annotators from other cultures in order to collect responses from different cultures for each template. From these annotations, the authors of BLEND constructed a multilingual dataset of questions with short answers in the languages of their own cultures. In addition, the questions and answers were translated into English to form a Multiple-Choice Question dataset (MCQ) as shown in Figure 1. In our study, we focus on this set of multiple-choice questions, which is only available in English.

3.1.2 Data preprocessing

Training and evaluation samples are drawn from the BLEND dataset and adapted to our task as follows. The BLEND data sequences include the question, the options, the instructions, and the answer (the letter of the correct answer among the options), as shown in Figure 1. We partition the data into training and evaluation sets, taking care to group questions of the same template together in or-

der to evaluate the model exclusively on unknown templates.

Let $\mathcal{D} = (Q_k)_{k\in\{1,\dots,N\}}$ be the set of question templates. We separate the templates used for training $\mathcal{D}_{train} = (Q_k)_{1\leq k\leq N_{train}}$ from the evaluation templates $\mathcal{D}_{test} = (Q_k)_{N_{train}+1\leq k\leq N}$. For each question template Q_k, the questions adapted to each culture c are denoted $Q_k(c)$ (for example, with the mention 'in Spain" for $c =$ Spain). The set of options $(a_j)_{j\in[A,B,C,D]}$ contains the answer to Q_k for culture c and those for three other cultures. Finally, as illustrated in Figure 1, the questions are accompanied by instructions I_1 and I_2:

$$q_k(c)_{(a_A,a_B,a_C,a_D)} =$$
$$Q_k(c) \oplus I_1 \oplus \left(\bigoplus_{j\in[A,B,C,D]} j \oplus \text{“.”} \oplus a_j \right) \oplus I_2$$

The original dataset required a response in JSON format. Since the results of the base model (see Section 4.1) were unsatisfactory, we modified the instruction so that the model would respond only with the letter corresponding to the answer (instruction I_1 in Figure 1). This instruction was retained for training and evaluation.

3.2 Parameter-efficient fine-tuning

We fine-tune our models on the next token prediction task, training them to predict the correct one-letter answer (*e.g.*, "A"). Let us consider the short sequence of tokens $y = (y_1, \dots, y_T)$[2] corresponding to the response to question $q_k(c)_{(a_A,a_B,a_C,a_D)}$, we note $y_{<t} = (y_1, \dots, y_{t-1})$. With p_θ being the conditional probability distribution of the model with parameters θ over the next token (θ includes the frozen parameters of the base model and the trainable parameters of the soft prompt), the calculated loss is as follows:[3]

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p_\theta \left(y_t \middle| q_k(c)_{(a_A,a_B,a_C,a_D)}, y_{<t} \right)$$

Prompt-Tuning The trainable parameters are a matrix $P \in \mathbb{R}^{n\times d}$ where n is the length of the soft prompt (the number of virtual tokens), and d is the dimension of the model's embedding.

[2] y being composed of the response proper (*e.g.*, "A") followed by the sequence of tokens marking the end of the model's response.

[3] Due to the limited amount of data we have, we use the same loss function as in the training of the generative model, rather than a classification loss function.

Prefix-Tuning The trainable parameters are a matrix $P \in \mathbb{R}^{n\times(2Ld)}$ where L is the number of layers. The prefix can be projected with a 2-layer multilayer perceptron of intermediate dimension p. The number of trainable parameters is then $p(d + 1) + 2Ld(p + 1) + nd$, but only the final prefix of size $n \times (2Ld)$ is saved and reused at inference.

3.3 Soft prompts interpretation methods

Soft prompts are easily manipulable to understand what was learned during training. While previous work stress on the difficulty to project continuous tokens on the discrete space (see Section 2.2), we suggest to interpret the cultural prompts by looking into their similarities.

Similarities between prompts of different cultures Intuitively, if two cultures are related by common history, linguistic or cultural characteristics, their corresponding trained prompts are more likely to be close in the embedding space. To study this hypothesis, we looked at the cosine similarities between soft prompts. In our experiments, we observe that soft prompts trained from a given initialization are still very close in the embedding space at the end of the training. To mitigate the influence of the initialization, we consider the concatenation of the prompts from different initializations to get a culture vector for each culture. Then, we apply a dimension reduction (Principal Component Analysis) and measure cosine similarities with the k first components of the culture vectors.

4 Experiments and results

4.1 Experimental setup

Data splits To enable model evaluation across cultures, we carefully split our training and evaluation datasets according to the question templates. We then extract the corresponding training and evaluation sets for each culture, resulting in different dataset sizes, as for each template, we obtain varying numbers of questions across cultures. To ensure a fair comparison, we select three data splits (see Appendix A.1) where evaluation proportions are the most similar for every culture (selecting the splits with the smallest difference between the maximum and the minimum evaluation proportion sizes).

Model fine-tuning We select the conversational model OLMo-7B-Instruct (Groeneveld et al., 2024)

	GB	US	ES	MX	GR	CN	ID	KR	IR	ET	AZ	DZ	NG	AS	JB	KP
Baseline	77.0	67.6	74.2	72.4	70.3	68.1	65.7	66.2	62.0	55.6	54.9	55.1	53.5	51.6	43.0	49.9
Prompt-t.	85.1	**85.8**	80.6	**80.6**	**79.5**	**81.4**	**79.4**	76.0	74.8	67.4	72.6	72.6	**72.5**	67.6	**69.3**	**69.1**
Prefix-t.	**87.0**	85.5	**82.0**	79.6	78.5	**81.4**	78.4	**76.2**	**78.6**	**72.0**	**74.2**	**74.1**	72.0	**71.3**	68.8	65.0
LoRA	86.9	82.4	79.8	79.6	78.6	79.2	77.3	73.0	74.6	70.4	70.4	71.4	66.0	65.7	65.0	62.0

Table 1: Proportion of correct answers of soft prompt models evaluated on their target culture. The best-performing results are highlighted in bold, while the second-best results are underlined. The country/region codes correspondences are the following. GB: United Kingdom; US: United States; ES: Spain; MX: Mexico; GR: Greece; CN: China; ID: Indonesia; KR: South Korea; IR: Iran; ET: Ethiopia; AZ: Azerbaijan; DZ: Algeria; NG: Northern Nigeria; AS: Assam; JB: West Java; KP: North Korea. More detailed results are provided in the Appendix A.2.

available on HuggingFace as our base model.[4] We fine-tune our models with the PEFT library[5] with a loss function computed only on the assistant's response. We train our models for 30 epochs, with batch sizes of 16. Prompt-tuning is performed with 16 virtual tokens and a learning rate of 10^{-4}. Prefix-tuning uses 4 virtual tokens with a projection size of 256 and a learning rate of 10^{-7}. As mentioned in Section 2.2, we compare our approaches with the widely used LoRA. LoRA rank was fixed at 2 with a learning rate of 10^{-7}. These hyperparameters were chosen based on preliminary experiments. We replicate our experiments six times, initializing our soft prompts with two different seeds for each of our three data splits. Our results are then averaged on these six experiments. For the baseline, we use the base model in a zero-shot setting, using the same question format and instructions as in the fine-tuning dataset, as shown in Figure 1.

Post-processing and evaluation metric We evaluate our models on BLEND's evaluation subset using the accuracy (the percentage of multiple-choice questions on which the model selects the correct letter). Early experiments show that the baseline struggles to answer only the letter (*e.g.* "B"), therefore we allow the letter to be provided along with its corresponding attribute (*e.g.* "B. tacos"). While the baseline sometimes outputs invalid answers, these are minor errors and are counted as incorrect. The PEFT models are capable of answering just the letter without any difficulty.

Evaluation settings We first investigate how the soft prompt models perform on the culture they were fine-tuned on, which we call in-culture evaluation. Then, we wonder how the fine-tuning on a specific culture influences the results on other cultures. We have two hypotheses for these cross-cultural evaluation. Our first hypothesis is that catastrophic forgetting occurs, causing the model to ignore the culture mentioned in the question and respond according to its fine-tuning culture. Our second hypothesis is that a model's performance on a culture will reflect the links between the fine-tuning culture and the evaluated one.

[4] https://huggingface.co/allenai/OLMo-7B-Instruct-hf. Note that this model is older than the BLEND corpus, thus avoiding contamination.

[5] https://huggingface.co/docs/peft/index

Method	Hyperp.	# trained	# saved
Prompt-t.	$n = 16$	65K (0.001 %)	65K (0.001 %)
Prefix-t.	$n = 4, p = 256$	68,000K (0.984 %)	1,000K (0.015 %)
LoRA	$r = 2$	1,000K (0.015 %)	1,000K (0.015 %)

Table 2: Number of trainable and saved parameters of PEFT methods. Notations: n is the number of virtual tokens, p is the projection dimension for prefix-tuning, and r corresponds to the LoRA rank.

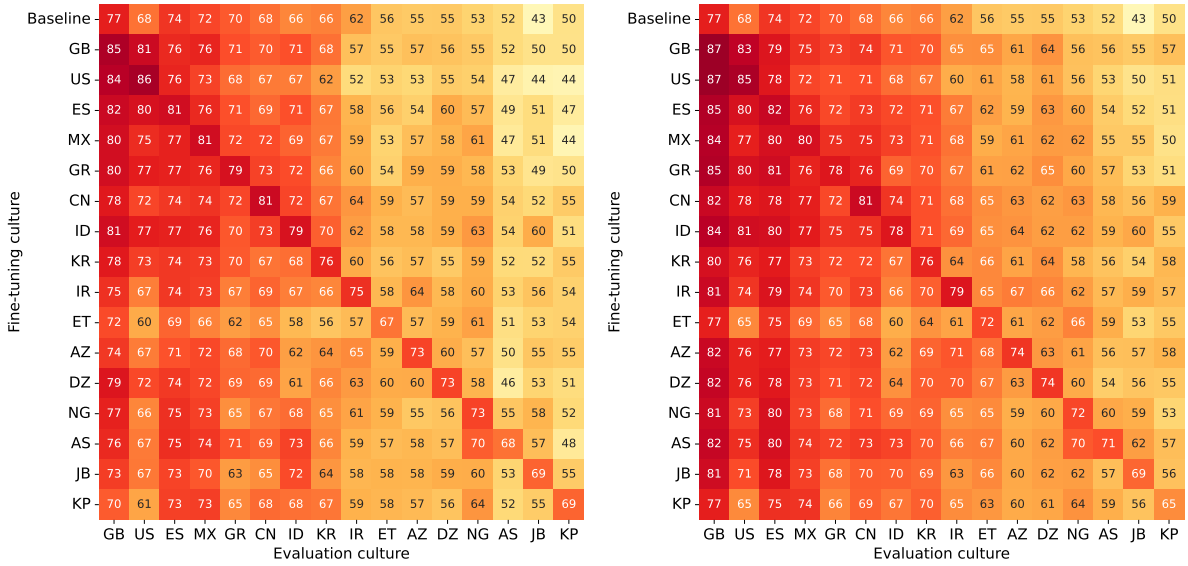
4.2 Results

4.2.1 Performance of soft prompt models

Baseline performance Table 1 shows the models' performance on the BLEND test subsets. We first observe disparate results of the baseline, indicating a lack of knowledge of some cultures, especially low-resource ones (*e.g.*, Assam, North Korea, and West Java). Surprisingly, the performance on the United States is low compared to other high-resource Western cultures (*e.g.*, the United Kingdom, Spain, and Mexico).

In-culture results A second key observation is that all PEFT methods improved results compared to the baseline. LoRA, as the main used PEFT method in the literature, underperforms soft prompts methods in our setting for all the cultures.

As the lighter method, prompt-tuning brings



(a) Prompt-tuning

(b) Prefix-tuning

Figure 2: Cross-culture results. Country/region codes are explained in the caption of Table 1. More detailed results are provided in the Appendix A.2.

gains from 6.4 (Spain) to 26.3 points (West Java) by training only 0.001% of the model parameters (Table 2). The larger margins are especially observable on low-resource cultures (*e.g.*, Northern Nigeria, Assam, West Java, and Azerbaijan).

Prefix-tuning, while training almost 1000 times more parameters than prompt-tuning, reaches similar results. Depending on the culture, it slightly outperforms prompt-tuning, especially for low-resource cultures such as Iran, Ethiopia, Algeria, Azerbaijan, and Assam, whereas it stays slightly behind prompt-tuning for Northern Nigeria and West Java. North Korea is the only low-resource culture where prompt-tuning outperforms prefix-tuning with a larger margin.

Appendix A.2 details the standard deviations of these accuracies, revealing noticeable differences across the experiments (differing by their data splits and soft prompt initializations), and highlighting no clear distinction between prefix-tuning and prompt-tuning, while both show significant improvement over the baseline.

Cross-culture results Figure 2 shows the cross-cultural results for soft prompt methods. Column-wise results indicate that almost all cross-cultural adaptation settings lead to improved performance compared to the baseline, thereby ruling out the catastrophic forgetting hypothesis. However, there are some exceptions. For instance, fine-tuning on Ethiopia reduces performance on high-resource cultures (*e.g.*, Spain, China, United Kingdom, South Korea) for both prompt- and prefix-tuning.

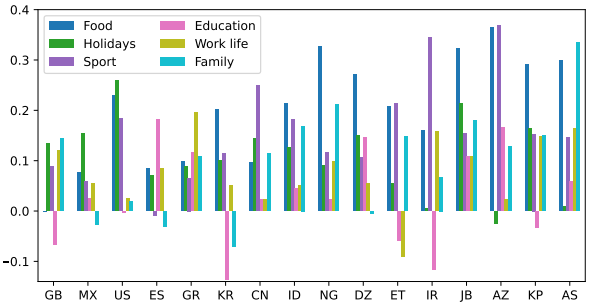
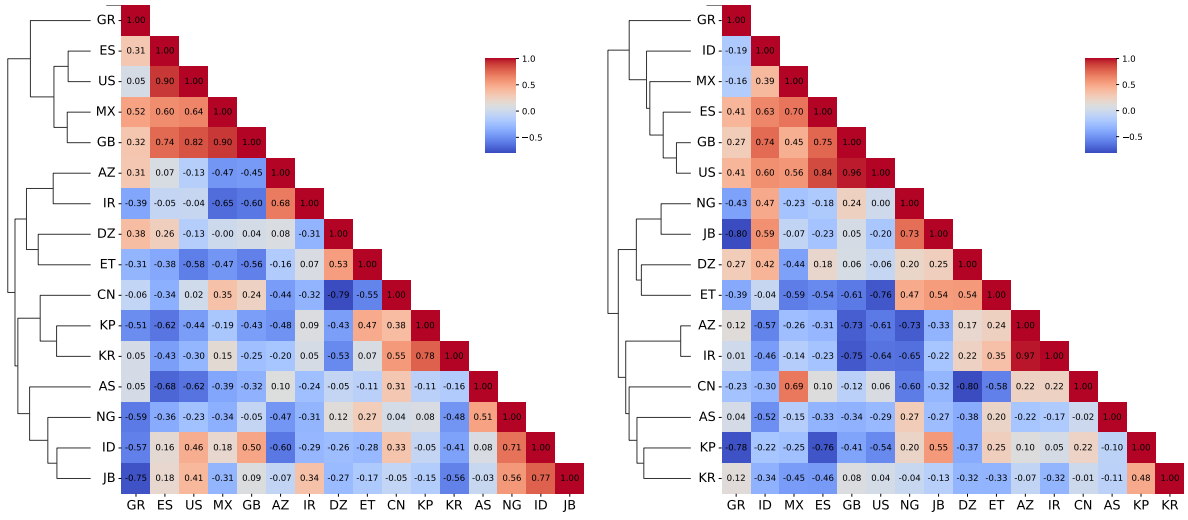


Figure 3: Performance gap between the baseline and prompt-tuned models across BLEND topics.

While in-culture evaluation shows that prefix-tuned models do not significantly outperform prompt-tuned models, we observe a consistent trend of improved cross-cultural performance with prefix-tuning. This suggests that prefix-tuning enables better generalization across cultural contexts, whereas prompt-tuning tends to be more specialized toward the target culture. Finally, we can observe some interesting insights about performance on other cultures. For instance, models fine-tuned on the United States show high results on the United Kingdom, even outperforming their target culture results in the case of prefix-tuning. From these results, we conclude that PEFT methods enable models to learn the MCQ format task



(a) Prompt-tuning (b) Prefix-tuning

Figure 4: Application of the PCA on the concatenation of the six trained soft prompts for each culture. Corresponding visualization of the two first dimensions is in Appendix C.

and cultural alignment outside their target culture [6]. However, while we observe a clear distinction between high- and low-resource cultures, we do not see clear patterns of similarities between cultures in these cross-cultural results, thus rendering it difficult to interpret cultural proximities.

Results per topic We hypothesize that performance varies across topics. Figure 3 shows the performance gains of prompt-tuned models over the baseline across cultures and BLEND topics. We observe that *Food* and *Sport* are the two topics that are best learned, with greater improvements on low-resource cultures, while, for *Education*, performance decreases slightly for six cultures. This could be explained by the repartition of the BLEND topics in our data splits (see Appendix A.3). Moreover, Appendix B shows the cross-cultural results on each BLEND topic, revealing that *Work life* and *Holidays/Celebration/Leisure* are the most easily learned topics across all cultures. Overall, cross-culture performance depends primarily on the relationship between the evaluated culture and topic, as illustrated by strong results for *Work life* on China and the United Kingdom, and for *Food* on Spain and Mexico.

[6]To verify that these cross-cultural capacities go beyond the MCQ format task, we conducted an ablation study in Appendix D.

4.2.2 Soft prompts in the embedding space

As detailed in the previous section (see Section 2.2), soft prompts are not directly comparable to hard prompts. However, following Vu et al. (2022), we explore the interpretation of soft prompts in the embedding space. We investigate whether cultures that share similarities have close prompts, leading to results that are easier to interpret than previous cross-culture results.

Similarities between soft prompts First, we investigate the links between prompts trained on different cultures. We follow our methodology described in Section 3.3 and we take $k = 5$ principal components to measure cosine similarities between culture vectors. Figure 4a shows the results for prompt-tuning. First, the culture clustering indicates a clear distinction between Western cultures and non-Western ones. In line with geo-cultural patterns, similarities between soft prompts are observed between Azerbaijan and Iran, between China and both North and South Korea, as well as between Indonesia and West Java. Moreover, low-resource non-Western cultures (*e.g.*, North Korea, Ethiopia, Assam) are shown to have negative cosine similarities with high-resource Western ones (*e.g.*, United Kingdom, United States, Spain). However, some similarity scores remain difficult to interpret intuitively (*e.g.*, Northern Nigeria close to Assam and Indonesia).

Prefix-tuning (Figure 4b) exhibits similar patterns. Interestingly, Indonesia is closer here to

Western cultures than to West Java, thus moving away from geographical similarities.

These results on prompt similarities do not really coincide with the cross-culture results discussed in Section 4.2.1. In fact, the cultural similarities derived from our soft prompts appear to be overall more easily interpretable than those obtained from the cross-culture evaluation presented in Figure 2.

5 Conclusion

In this paper, we investigate novel methods for aligning models on cultural commonsense knowledge at low cost. We use soft prompt methods (prompt-tuning and prefix-tuning) to compare with the commonly used LoRA, to fine-tune culture-specific models on the MCQ dataset of BLEnD on 16 cultures. Soft prompt methods significantly improve cultural alignment over the baseline, and outperform LoRA on average in our setting. We also show that models fine-tuned with soft prompts on a given culture often outperform the baseline when evaluated on another culture. These cross-cultural results reveal that all models perform better on high-resource and Western cultures. We also examine the ability of soft prompt approaches to interpret relationships between cultures present in BLEnD. We observe that the soft prompts capture geographical similarities (*e.g.*, Azerbaijan and Iran) as well as cultural ones (*e.g.*, the United Kingdom and the United States). In future work, we plan to extend our study of these soft prompt methods in other, more general culturally dependent settings (*e.g.*, longer, natural language answers).

Limitations

Our study is based on the BLEnD dataset, which was built on human annotations. Ethical considerations were raised by Myung et al. (2024) about the lack of representativeness of a few annotators for a whole region or country, which also applies to our study. We use the MCQ part of BLEnD data, but we are aware that aligning models with MCQ data does not take into account the full complexity of cultural commonsense knowledge, and does not measure the capacity of the model to use that knowledge in real-world situations.

Moreover, all our data is in English, which fails to capture the cultural vocabulary of the regional language and the important local specialities that could be lost in translation. However, it enables us to evaluate models across cultures fairly without the fluctuations of the models' multilingual capacities. Finally, we observed very disparate performances across the data splits for the baseline and the soft prompt models, underscoring the critical role of alignment and evaluation data selection.

Our study is limited to a single base model (OLMo-7B-Instruct). In our experiments, we focus on studying cultural similarities, which requires training and evaluation on the 16 cultures of BLEnD. To limit the number of experiments, we had to restrict our study to a single base model. The model choice was oriented by the fact that its date cutoff is prior to the release of BLEnD, ensuring that the model was not trained on BLEnD.

Ethical considerations

Our study focuses on culture, a concept that is difficult to defined and to deal with, in particular in NLP. As discussed in Section 2, we made the choice to use cultural commonsense knowledge as a proxy that is generalisable across a country or a region and static over time. However, we acknowledge that culture is a more nuanced concept that extends beyond commonsense knowledge and evolves over time. A more global study of soft prompts for cultural adaptation could include alignment with other proxies, such as real-world situations to apply cultural values and norms. We also emphasize that cultural adaptation raises important ethical concerns. Excessive adaptation may reinforce polarized views (Rao et al., 2025) and contribute to the propagation of cultural stereotypes and demographic profiling (Adilazuarda et al., 2024). Finally, we acknowledge that the interpretation of the results is not exempt from the authors' biases.

Acknowledgments

This work was partly funded by the ANR project SINNET (ANR-23-CE23-0033). It was also partly funded by Benoît Sagot's chairs in the PRAIRIE institute, funded by the French national agency ANR, as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and by Benoît Sagot's chair in its follow-up, PRAIRIE-PSAI, also funded by the ANR as part of the "France 2030" strategy under the reference ANR23-IACL-0008. The authors are grateful to the CLEPS infrastructure from the Inria Paris for providing resources and support. This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011016786 made by

GENCI.

References

Farid Adilazuarda, Chen Cecilia Liu, Iryna Gurevych, and Alham Fikri Aji. 2025. From Surveys to Narratives: Rethinking Cultural Value Adaptation in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18052–18079, Suzhou, China. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From Word to World: Evaluate and Mitigate Culture Bias in LLMs via Word Association Test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24510–24526, Suzhou, China. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *Preprint*, arXiv:2306.16388.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier L. de Lacalle, and Mikel Artetxe. 2024. BertaQA: How Much Do Language Models Know About Local Culture? *Advances in Neural Information Processing Systems*, 37:34077–34097.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A Fine-grained Cultural-aware LLM Training Paradigm via Multilingual Critique Data Synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria. Association for Computational Linguistics.

Clifford Geertz. 1973. *The Interpretation of Cultures*. Basic Books.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual Human Preference Learning for Cultural Awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32854–32883, Suzhou, China. Association for Computational Linguistics.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey: Round seven – country-pooled datafile version 6.0.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *The Behavioral and Brain Sciences*, 33(2-3):61–83; discussion 83–135.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Tianyi Hu, Maria Maistro, and Daniel Hershcovich. 2024. Bridging Cultures in the Kitchen: A Framework and Benchmark for Cross-Cultural Recipe Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1080, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. 2022. Prompt Waywardness: The Curious Case of Discretized Interpretation of Continuous Prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3631–3643, Seattle, United States. Association for Computational Linguistics.

Piyawat Lertvittayakumjorn, David Kinney, Vinodkumar Prabhakaran, Donald Martin Jr., and Sunipa Dev.

2025. Towards Geo-Culturally Grounded LLM Generations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–330, Vienna, Austria. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating Cultural Differences into Large Language Models. In *NeurIPS 2024*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025b. Cultural Learning-Based Culture Adaptation of Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Chunhua Liu, Kabir Manandhar Shrestha, and Sukai Huang. 2025c. ALIGN: Word Association Learning for Cross-Cultural Generalization in Large Language Models. *Preprint*, arXiv:2508.13426.

H. Liu and P. Singh. 2004. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 22(4):211–226.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2024. GPT understands, too. *AI Open*, 5:208–215.

Reem I. Masoud, Martin Ferianc, Philip Colin Treleaven, and Miguel R. D. Rodrigues. 2024. LLM Alignment Using Soft Prompt Tuning: The Case of Cultural Alignment. In *Workshop on Socially Responsible Language Modelling Research*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki A. Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew A. Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen H. Muhammad, Kiwoong Park, Anar S. Rzayev, Nina White, Seid M. Yimam, Mohammad T. Pilehvar, and 3 others. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting Cultural Commonsense Knowledge at Scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pages 1907–1917, New York, NY, USA. Association for Computing Machinery.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural Commonsense Knowledge for Intercultural Dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 1774–1784, New York, NY, USA. Association for Computing Machinery.

Eric J. W. Orlowski, Hakim Norhashim, and Tristan Koh Ly Wey. 2025. 'Too much alignment; not enough culture': Re-balancing cultural alignment practices in LLMs. *Preprint*, arXiv:2509.26167.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025a. Survey of Cultural Awareness in Language Models: Text and Beyond. *Computational Linguistics*, pages 1–96.

Siddhesh Milind Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025b. Presumed Cultural Identity: How Names Shape LLM Responses. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22147–22172, Suzhou, China. Association for Computational Linguistics.

Pew Research Center. 2026. Pew Global Attitudes Survey. `https://www.pewresearch.org`.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM Generate Culturally Relevant Commonsense QA Data? Case Study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. Association for Computational Linguistics.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable Natural Language Generation with Contrastive Prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924, Dublin, Ireland. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models. In *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 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Ao Sun, Xiaoyu Wang, Zhe Tan, Yu Li, Jiachen Zhu, Shu Su, and Yuheng Jia. 2026. CuMA: Aligning LLMs with Sparse Cultural Values via Demographic-Aware Mixture of Adapters. *Preprint*, arXiv:2601.04885.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Edward B. Tylor. 1871. *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art and Custom*. John Murray, London.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022. SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland. Association for Computational Linguistics.

Haonan Wang, Brian K. Chen, Li Siquan, Liang Xinhe, Tianyang Hu, Hwee Kuan Lee, and Kenji Kawaguchi. 2025. Prefix-Tuning+: Modernizing Prefix-Tuning by Decoupling the Prefix from Attention. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! At ICML 2025*.

Yifan Wang and Vera Demberg. 2024. A Parameter-Efficient Multi-Objective Approach to Mitigate Stereotypical Bias in Language Models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 1–19, Bangkok, Thailand. Association for Computational Linguistics.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023. Tailor: A Soft-Prompt-Based Approach to Attribute-Based Controlled Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–427, Toronto, Canada. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, Jindong Wang, Zhicheng Dou, and Xing Xie. 2025. CAReDiO: Cultural Alignment of LLM via Representativeness and Distinctiveness Guided Data Optimization. *Preprint*, arXiv:2504.08820.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, Yuxiao Dong, and Jie Tang. 2025. Parameter-Efficient Fine-Tuning for Foundation Models. *Preprint*, arXiv:2501.13787.

Zhen-Ru Zhang, Chuanqi Tan, Haiyang Xu, Chengyu Wang, Jun Huang, and Songfang Huang. 2023. Towards Adaptive Prefix Tuning for Parameter-Efficient Language Model Fine-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1239–1248, Toronto, Canada. Association for Computational Linguistics.

A Data details

A.1 Data splits

Figure 5 details the number of individual questions used to train and evaluate our models on each culture, for the three selected data splits (seeds).

A.2 Detailed results

Baseline We detail the mean accuracy and standard deviation of the baseline on the evaluation sets of each country in Table 3.

Prompt-tuning Tables 4, 5, 6, and 7 show the detailed accuracies and standard deviations of the prompt-tuned models across fine-tuning countries and evaluation countries.

Prefix-tuning Similar results for prefix-tuning are shown in Tables 8, 9, 10, and 11.

A.3 Topics distributions

Figures 6, 7, and 8 show the distributions of the BLEND topics over training and evaluation samples, by data split (seed) and culture.

B Cross-cultural results across topics

The cross-cultural performance of prompt-tuned models is presented by topics in Figure 9.

Evaluation country	Accuracy	Standard deviation
Algeria	55.08	7.01
Assam	51.61	5.43
Azerbaijan	54.90	1.37
China	68.06	3.34
Ethiopia	55.62	5.15
Greece	70.25	7.32
Indonesia	5.66	7.76
Iran	62.03	2.31
Mexico	72.37	1.20
North Korea	49.86	9.70
Northern Nigeria	53.47	9.25
South Korea	66.20	5.06
Spain	74.24	8.57
UK	76.96	4.48
US	67.61	7.81
West Java	42.97	9.38

Table 3: Baseline — Detailed results.

C Visualization of the soft prompts

Alongside Figure 4, which presents cosine similarities between culture vectors computed on the first five PCA dimensions, Figure 10 visualizes the projection of these vectors onto the first two principal components.

D Ablation study: Prompt-tuning on the MCQ format task

Cross-culture results (see Section 4.2.1) indicate that fine-tuning on a culture improve the model's performances on other cultures. We wonder if this observation might be due to the fact that the model is fine-tuned on the MCQ format (i.e., answering "A", "B", "C" or "D"). To test this hypothesis, we fine-tuned a model on the MCQ task without any mention of a specific culture.

Dataset creation We kept the previous data splits of the BLEND templates. We extract the corresponding questions for all cultures and preprocess the questions by removing any mention of a culture (e.g., removing "in Spain", "of Mexico"). We randomly sample among these questions to obtain training and test sets of similar sizes to those used in previous experiments. We ensure that the correct answers correspond to the different cultures in a balanced way.

Fine-tuning We conducted this study on prompt-tuning, keeping the same hyperparameters as in previous experiments. The experiments were replicated 6 times, on which we average the results.

Results The average accuracy obtained on the test set is 28.33% (close to 25%), indicating the ability to learn to answer one of the four options, almost without no preference between them. Moreover, Figure 11 shows the accuracies on the test sets of the 16 cultures. The performances are lower than the baseline in a zero-shot setting used in our main experiments. We conclude that prompt-tuning on the MCQ format task does not improve the performances over our baseline and therefore does not outperform the models fine-tuned on the other cultures. This evaluation shows similar trends to our baseline, favoring high-resource over low-resource cultures. The transfer capacities of the models fine-tuned on a country and evaluated on another one are therefore not due to the MCQ format.

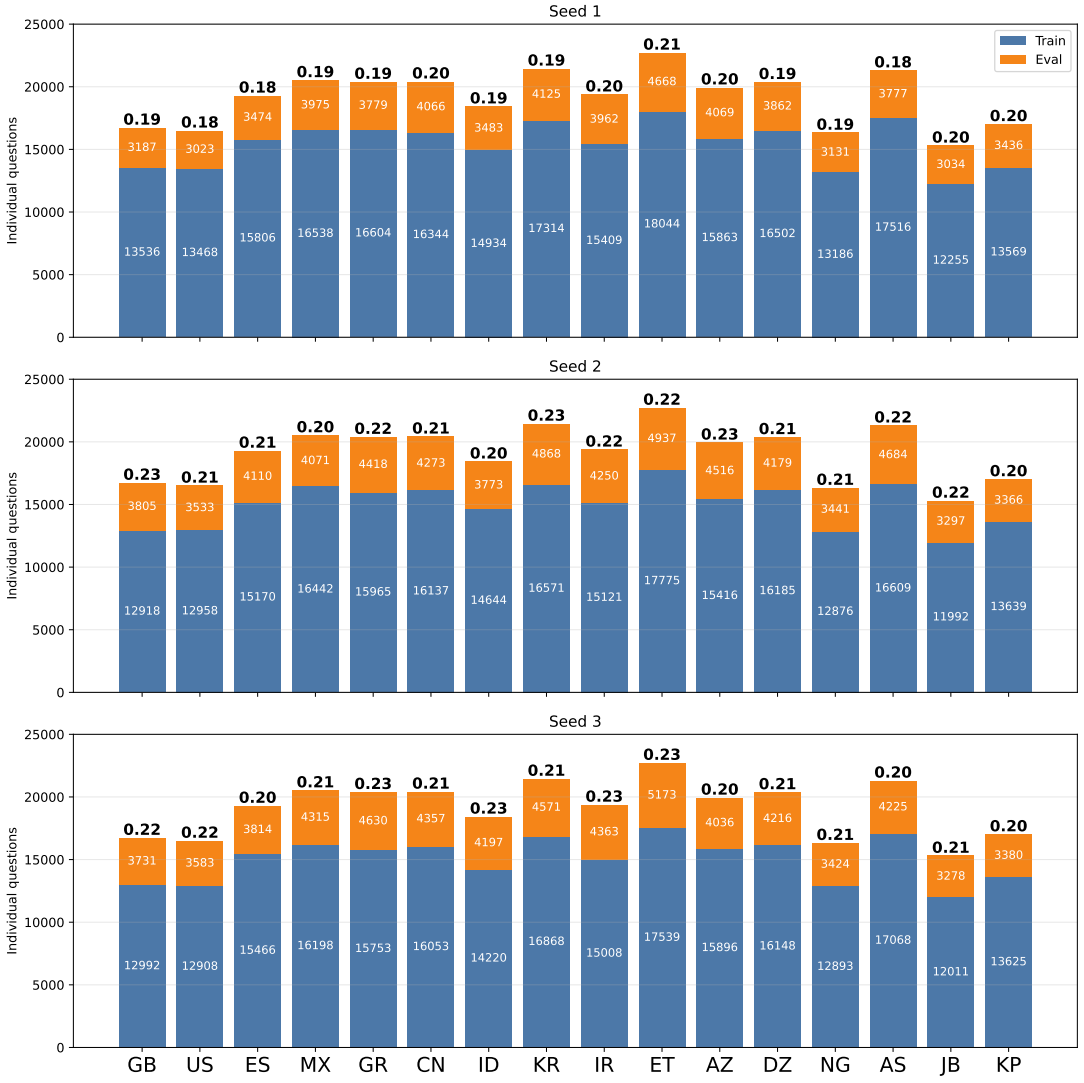


Figure 5: Data splits — Distributions between train and evaluation sets of the individual questions for each culture. The proportion of evaluation data is indicated in bold above the bars.

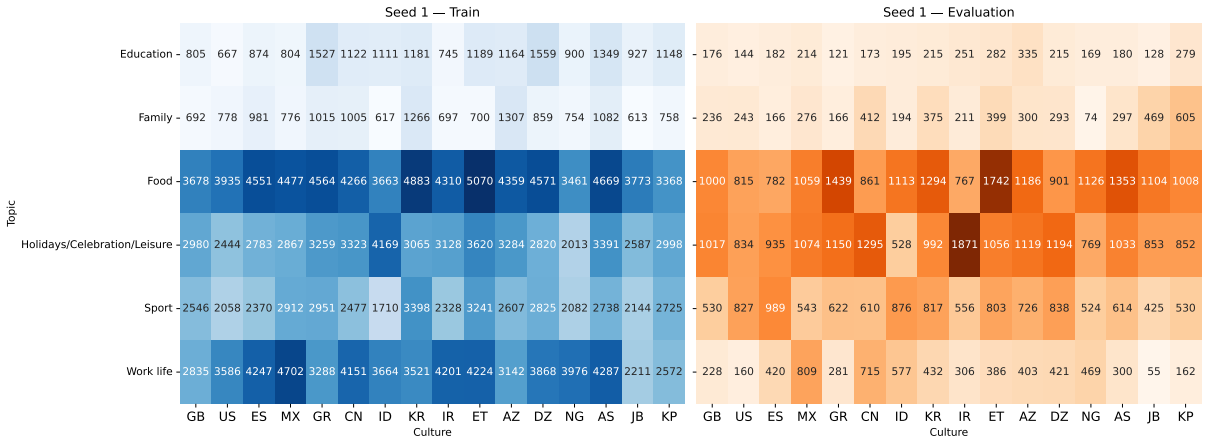


Figure 6: Topics distributions of training and evaluation questions in the data split of seed 1.

Training country	Evaluation country	Accuracy	Standard deviation
Algeria	Algeria	72.57	5.32
	Assam	46.44	7.98
	Azerbaijan	60.05	3.18
	China	68.51	4.98
	Ethiopia	59.75	5.02
	Greece	68.79	6.53
	Indonesia	61.14	9.54
	Iran	62.83	8.75
	Mexico	72.18	8.75
	North Korea	51.50	5.39
	Northern Nigeria	58.33	6.86
	South Korea	65.69	5.64
	Spain	73.68	11.51
	UK	78.52	6.58
	US	71.94	8.81
	West Java	53.18	10.72
Assam	Algeria	56.90	5.92
	Assam	67.62	3.84
	Azerbaijan	57.91	5.02
	China	69.47	5.53
	Ethiopia	57.21	3.35
	Greece	70.80	7.96
	Indonesia	72.70	11.39
	Iran	58.59	5.74
	Mexico	74.12	8.25
	North Korea	47.56	5.72
	Northern Nigeria	69.87	5.72
	South Korea	65.63	2.97
	Spain	75.00	8.81
	UK	75.59	5.11
	US	66.80	7.19
	West Java	56.97	3.11
Azerbaijan	Algeria	59.57	8.70
	Assam	49.77	8.73
	Azerbaijan	72.56	5.52
	China	70.45	6.17
	Ethiopia	58.97	8.07
	Greece	68.37	6.59
	Indonesia	61.94	9.54
	Iran	65.36	4.97
	Mexico	71.91	9.46
	North Korea	54.62	6.15
	Northern Nigeria	56.79	4.97
	South Korea	63.67	5.33
	Spain	70.89	11.13
	UK	73.87	7.19
	US	66.65	6.65
	West Java	55.44	6.72
China	Algeria	58.97	6.56
	Assam	53.76	5.84
	Azerbaijan	57.37	3.24
	China	81.37	3.57
	Ethiopia	58.82	3.95
	Greece	72.25	5.67
	Indonesia	72.24	8.25
	Iran	63.59	4.32
	Mexico	74.48	6.21
	North Korea	54.76	7.25
	Northern Nigeria	58.98	6.77
	South Korea	67.39	5.17
	Spain	73.88	5.40
	UK	77.89	3.97
	US	72.24	5.12
	West Java	52.13	4.57

Table 4: Prompt-tuning — Detailed results of models trained on Algeria, Assam, Azerbaijan, China.

Training country	Evaluation country	Accuracy	Standard deviation
Ethiopia	Algeria	58.69	7.82
	Assam	50.69	7.92
	Azerbaijan	57.31	5.23
	China	64.67	5.12
	Ethiopia	67.41	5.74
	Greece	62.42	4.15
	Indonesia	57.60	10.98
	Iran	57.27	7.93
	Mexico	66.01	7.97
	North Korea	53.72	5.88
	Northern Nigeria	61.36	6.80
	South Korea	55.68	4.75
	Spain	69.00	10.26
	UK	71.75	6.29
	US	59.99	9.39
	West Java	53.35	3.15
Greece	Algeria	59.30	8.15
	Assam	52.73	7.29
	Azerbaijan	59.06	2.55
	China	72.75	3.18
	Ethiopia	53.82	3.86
	Greece	79.47	7.40
	Indonesia	71.52	9.50
	Iran	59.51	3.51
	Mexico	76.12	6.92
	North Korea	49.53	7.63
	Northern Nigeria	57.87	7.90
	South Korea	65.96	3.17
	Spain	76.63	7.13
	UK	80.28	5.18
	US	77.13	9.95
	West Java	48.95	7.35
Indonesia	Algeria	58.98	8.38
	Assam	53.60	6.30
	Azerbaijan	57.91	5.37
	China	73.01	3.31
	Ethiopia	57.92	2.90
	Greece	70.37	6.83
	Indonesia	79.42	9.45
	Iran	62.27	3.51
	Mexico	76.30	6.21
	North Korea	50.97	5.98
	Northern Nigeria	62.97	6.12
	South Korea	69.65	6.48
	Spain	76.71	6.63
	UK	80.51	2.18
	US	77.27	7.09
	West Java	59.70	4.10
Iran	Algeria	57.69	8.17
	Assam	53.49	9.23
	Azerbaijan	63.74	4.67
	China	69.04	5.53
	Ethiopia	57.81	3.70
	Greece	66.51	6.84
	Indonesia	66.54	9.78
	Iran	74.83	5.09
	Mexico	73.15	7.82
	North Korea	53.84	9.95
	Northern Nigeria	59.69	7.55
	South Korea	65.62	5.51
	Spain	73.84	9.17
	UK	74.69	6.28
	US	66.53	7.90
	West Java	55.99	6.13

Table 5: Prompt-tuning — Detailed results of models trained on Ethiopia, Greece, Indonesia, Iran.

Training country	Evaluation country	Accuracy	Standard deviation
Mexico	Algeria	58.34	8.37
	Assam	47.28	6.11
	Azerbaijan	56.96	4.09
	China	71.52	3.57
	Ethiopia	53.36	9.55
	Greece	72.01	9.11
	Indonesia	69.12	8.36
	Iran	59.13	4.29
	Mexico	80.59	5.92
	North Korea	43.53	6.29
	Northern Nigeria	60.99	7.39
	South Korea	66.88	5.09
	Spain	76.64	7.27
	UK	80.41	4.85
	US	75.27	9.88
	West Java	51.25	6.01
North Korea	Algeria	55.98	8.18
	Assam	52.12	6.72
	Azerbaijan	57.41	4.87
	China	68.35	4.79
	Ethiopia	58.12	4.67
	Greece	65.47	3.78
	Indonesia	67.92	11.58
	Iran	58.79	3.04
	Mexico	73.09	11.30
	North Korea	69.10	4.00
	Northern Nigeria	63.76	8.18
	South Korea	67.27	6.35
	Spain	72.84	10.90
	UK	69.96	6.47
	US	61.04	9.97
	West Java	54.61	4.65
Northern Nigeria	Algeria	56.09	8.57
	Assam	55.43	7.32
	Azerbaijan	55.36	5.23
	China	67.15	6.95
	Ethiopia	58.69	7.30
	Greece	64.71	7.46
	Indonesia	68.31	11.55
	Iran	61.17	7.11
	Mexico	72.95	10.24
	North Korea	52.20	6.48
	Northern Nigeria	72.54	3.77
	South Korea	64.64	4.77
	Spain	74.95	9.58
	UK	76.69	3.84
	US	65.71	8.33
	West Java	58.24	7.03
South Korea	Algeria	54.96	8.60
	Assam	52.04	5.69
	Azerbaijan	57.03	5.74
	China	67.39	3.35
	Ethiopia	56.13	4.04
	Greece	70.15	8.93
	Indonesia	67.98	10.17
	Iran	59.88	3.94
	Mexico	73.42	8.26
	North Korea	54.97	4.99
	Northern Nigeria	59.27	6.80
	South Korea	75.95	4.66
	Spain	73.97	9.64
	UK	77.75	6.48
	US	73.16	5.00
	West Java	51.88	4.92

Table 6: Prompt-tuning — Detailed results of models trained on Mexico, North Korea, Northern Nigeria, South Korea.

Training country	Evaluation country	Accuracy	Standard deviation
Spain	Algeria	59.90	6.00
	Assam	48.54	7.18
	Azerbaijan	54.04	3.86
	China	68.65	3.51
	Ethiopia	55.52	2.23
	Greece	71.11	5.30
	Indonesia	71.15	8.19
	Iran	58.24	4.35
	Mexico	75.73	6.92
	North Korea	46.70	7.05
	Northern Nigeria	57.29	5.51
	South Korea	67.11	4.69
	Spain	80.58	6.06
	UK	81.50	4.06
	US	79.60	8.26
	West Java	51.40	8.07
UK	Algeria	56.20	9.10
	Assam	51.56	5.23
	Azerbaijan	56.52	6.59
	China	70.05	3.93
	Ethiopia	55.08	3.72
	Greece	71.37	6.19
	Indonesia	71.12	10.47
	Iran	56.53	3.87
	Mexico	76.18	7.16
	North Korea	49.71	7.79
	Northern Nigeria	55.11	8.92
	South Korea	67.83	4.33
	Spain	75.59	7.71
	UK	85.13	4.60
	US	81.40	6.48
	West Java	49.67	9.35
US	Algeria	54.63	8.16
	Assam	46.64	5.93
	Azerbaijan	52.65	5.69
	China	66.88	3.90
	Ethiopia	53.43	4.69
	Greece	68.24	6.38
	Indonesia	66.69	9.86
	Iran	51.67	3.86
	Mexico	72.62	6.42
	North Korea	43.83	6.02
	Northern Nigeria	54.00	7.05
	South Korea	62.35	6.67
	Spain	76.23	8.05
	UK	84.04	4.81
	US	85.82	5.89
	West Java	44.37	7.65
West Java	Algeria	59.41	10.89
	Assam	53.20	6.49
	Azerbaijan	58.26	5.16
	China	64.92	5.28
	Ethiopia	58.43	5.07
	Greece	62.52	6.57
	Indonesia	71.78	11.59
	Iran	58.08	7.10
	Mexico	70.43	9.20
	North Korea	54.68	6.21
	Northern Nigeria	59.72	8.98
	South Korea	63.62	5.72
	Spain	72.57	5.91
	UK	73.28	5.97
	US	67.00	7.05
	West Java	69.26	3.47

Table 7: Prompt-tuning — Detailed results of models trained on Spain, UK, US, West Java.

Training country	Evaluation country	Accuracy	Standard deviation
Algeria	Algeria	74.12	5.41
	Assam	54.43	6.48
	Azerbaijan	62.56	3.21
	China	72.15	4.37
	Ethiopia	66.85	4.51
	Greece	71.44	6.56
	Indonesia	64.29	5.48
	Iran	69.70	7.77
	Mexico	73.15	9.08
	North Korea	55.35	7.44
	Northern Nigeria	60.11	7.96
	South Korea	69.73	5.99
	Spain	77.98	9.38
	UK	82.41	5.81
	US	76.44	4.00
	West Java	55.65	7.04
Assam	Algeria	62.23	7.78
	Assam	71.31	3.93
	Azerbaijan	60.19	3.59
	China	73.41	4.41
	Ethiopia	66.73	4.78
	Greece	72.47	7.69
	Indonesia	72.59	9.62
	Iran	66.47	3.90
	Mexico	73.80	8.50
	North Korea	56.81	9.55
	Northern Nigeria	70.25	5.85
	South Korea	69.94	4.90
	Spain	79.81	7.61
	UK	82.29	5.31
	US	75.11	7.76
	West Java	62.10	4.85
Azerbaijan	Algeria	63.18	6.54
	Assam	56.46	4.30
	Azerbaijan	74.24	5.50
	China	72.95	5.43
	Ethiopia	67.76	5.23
	Greece	72.14	5.69
	Indonesia	62.29	8.31
	Iran	70.66	2.74
	Mexico	73.42	10.22
	North Korea	58.32	6.04
	Northern Nigeria	61.28	6.19
	South Korea	69.00	5.64
	Spain	77.14	11.14
	UK	81.61	4.85
	US	76.46	5.07
	West Java	56.75	4.20
China	Algeria	61.96	9.64
	Assam	57.90	5.33
	Azerbaijan	63.37	2.02
	China	81.41	3.92
	Ethiopia	65.09	5.23
	Greece	72.03	8.05
	Indonesia	74.36	9.55
	Iran	67.82	5.61
	Mexico	77.08	7.35
	North Korea	58.78	10.67
	Northern Nigeria	63.25	6.29
	South Korea	71.23	2.89
	Spain	78.34	6.97
	UK	81.60	4.94
	US	77.57	7.87
	West Java	56.46	6.66

Table 8: Prefix-tuning — Detailed results of models trained on Algeria, Assam, Azerbaijan, China.

Training country	Evaluation country	Accuracy	Standard deviation
Ethiopia	Algeria	62.29	7.31
	Assam	59.28	4.39
	Azerbaijan	60.77	3.86
	China	67.73	8.20
	Ethiopia	71.99	3.65
	Greece	65.28	7.61
	Indonesia	59.50	7.28
	Iran	61.13	9.60
	Mexico	68.62	9.46
	North Korea	55.38	7.75
	Northern Nigeria	66.03	8.05
	South Korea	63.92	6.55
	Spain	74.70	9.88
	UK	76.79	5.34
	US	65.45	9.19
	West Java	52.93	3.94
Greece	Algeria	64.91	7.69
	Assam	57.18	5.17
	Azerbaijan	61.53	1.76
	China	76.27	2.52
	Ethiopia	61.30	3.71
	Greece	78.46	7.88
	Indonesia	69.23	8.87
	Iran	66.93	2.01
	Mexico	76.35	7.03
	North Korea	51.05	7.07
	Northern Nigeria	60.30	5.14
	South Korea	70.31	3.15
	Spain	81.00	5.39
	UK	84.76	5.10
	US	80.02	7.40
	West Java	53.02	4.52
Indonesia	Algeria	62.22	7.44
	Assam	58.52	4.16
	Azerbaijan	63.85	4.16
	China	74.60	3.16
	Ethiopia	64.64	4.28
	Greece	74.53	7.87
	Indonesia	78.43	12.17
	Iran	68.59	4.36
	Mexico	77.32	8.01
	North Korea	55.15	7.13
	Northern Nigeria	62.15	5.46
	South Korea	71.23	4.05
	Spain	80.09	6.54
	UK	83.92	2.59
	US	80.57	4.89
	West Java	60.49	5.56
Iran	Algeria	65.74	6.72
	Assam	57.47	5.49
	Azerbaijan	66.59	3.74
	China	72.87	4.31
	Ethiopia	65.41	4.19
	Greece	70.22	6.31
	Indonesia	66.06	8.38
	Iran	78.59	5.30
	Mexico	74.21	9.14
	North Korea	57.36	7.71
	Northern Nigeria	62.24	6.88
	South Korea	69.67	4.30
	Spain	78.89	8.47
	UK	81.44	5.99
	US	74.19	6.08
	West Java	59.29	5.68

Table 9: Prefix-tuning — Detailed results of models trained on Ethiopia, Greece, Indonesia, Iran.

Training country	Evaluation country	Accuracy	Standard deviation
Mexico	Algeria	62.48	7.44
	Assam	55.11	6.20
	Azerbaijan	61.07	2.78
	China	75.00	4.12
	Ethiopia	59.34	5.67
	Greece	74.95	7.41
	Indonesia	72.86	9.51
	Iran	67.69	4.99
	Mexico	79.65	6.81
	North Korea	49.89	8.45
	Northern Nigeria	62.32	5.16
	South Korea	70.99	3.54
	Spain	79.70	4.91
	UK	84.40	4.83
	US	76.86	9.31
	West Java	55.16	5.87
North Korea	Algeria	60.57	8.36
	Assam	59.17	4.87
	Azerbaijan	59.95	3.52
	China	69.48	4.63
	Ethiopia	63.42	4.28
	Greece	65.78	6.56
	Indonesia	67.27	12.76
	Iran	65.16	6.23
	Mexico	74.41	10.62
	North Korea	65.04	5.68
	Northern Nigeria	64.15	7.15
	South Korea	69.68	5.94
	Spain	74.88	8.16
	UK	77.08	3.79
	US	65.33	7.31
	West Java	56.07	7.06
Northern Nigeria	Algeria	60.27	10.58
	Assam	60.34	3.86
	Azerbaijan	59.17	4.04
	China	71.31	5.72
	Ethiopia	65.31	5.99
	Greece	68.21	6.96
	Indonesia	68.61	10.25
	Iran	64.70	7.94
	Mexico	72.75	9.73
	North Korea	52.97	7.58
	Northern Nigeria	72.05	4.60
	South Korea	69.18	5.64
	Spain	79.57	5.62
	UK	81.29	4.15
	US	73.15	9.13
	West Java	59.26	4.54
South Korea	Algeria	64.47	7.53
	Assam	56.20	3.95
	Azerbaijan	61.13	3.53
	China	71.56	2.26
	Ethiopia	65.50	2.55
	Greece	71.67	7.84
	Indonesia	67.37	7.94
	Iran	64.35	5.83
	Mexico	73.39	7.78
	North Korea	57.87	5.11
	Northern Nigeria	58.04	4.96
	South Korea	76.19	3.19
	Spain	77.31	7.24
	UK	79.65	4.18
	US	76.14	5.08
	West Java	53.99	5.79

Table 10: Prefix-tuning — Detailed results of models trained on Mexico, North Korea, Northern Nigeria and South Korea.

Training country	Evaluation country	Accuracy	Standard deviation
Spain	Algeria	63.08	7.26
	Assam	54.49	6.77
	Azerbaijan	58.53	1.86
	China	73.47	2.87
	Ethiopia	61.91	5.23
	Greece	72.01	6.83
	Indonesia	71.87	8.92
	Iran	66.68	6.61
	Mexico	75.65	8.20
	North Korea	50.71	5.79
	Northern Nigeria	60.19	7.81
	South Korea	71.20	2.50
	Spain	82.03	6.18
	UK	84.82	3.03
	US	79.55	7.14
	West Java	51.97	7.67
UK	Algeria	63.94	9.83
	Assam	56.07	4.80
	Azerbaijan	60.52	3.75
	China	73.91	3.16
	Ethiopia	64.83	3.63
	Greece	73.04	6.14
	Indonesia	70.56	10.05
	Iran	64.78	6.54
	Mexico	74.70	9.24
	North Korea	56.51	7.91
	Northern Nigeria	56.50	6.52
	South Korea	70.26	3.70
	Spain	78.57	5.48
	UK	86.99	3.68
	US	83.30	5.33
	West Java	54.88	8.35
US	Algeria	60.76	10.10
	Assam	52.93	4.88
	Azerbaijan	58.19	4.70
	China	71.30	3.74
	Ethiopia	60.59	3.88
	Greece	71.23	5.52
	Indonesia	68.29	8.68
	Iran	59.63	4.47
	Mexico	72.41	9.22
	North Korea	50.93	6.56
	Northern Nigeria	56.25	6.70
	South Korea	67.05	4.09
	Spain	78.13	5.14
	UK	87.00	4.27
	US	85.47	5.39
	West Java	49.55	7.18
West Java	Algeria	61.83	8.54
	Assam	56.93	4.88
	Azerbaijan	60.31	4.88
	China	69.78	3.73
	Ethiopia	65.51	2.94
	Greece	67.56	7.25
	Indonesia	69.64	11.15
	Iran	63.30	4.92
	Mexico	72.87	7.89
	North Korea	56.28	10.18
	Northern Nigeria	62.25	6.40
	South Korea	68.55	4.95
	Spain	78.35	6.16
	UK	80.98	3.19
	US	71.48	4.98
	West Java	68.76	5.79

Table 11: Prefix-tuning — Detailed results of models trained on Spain, UK, US and West Java.

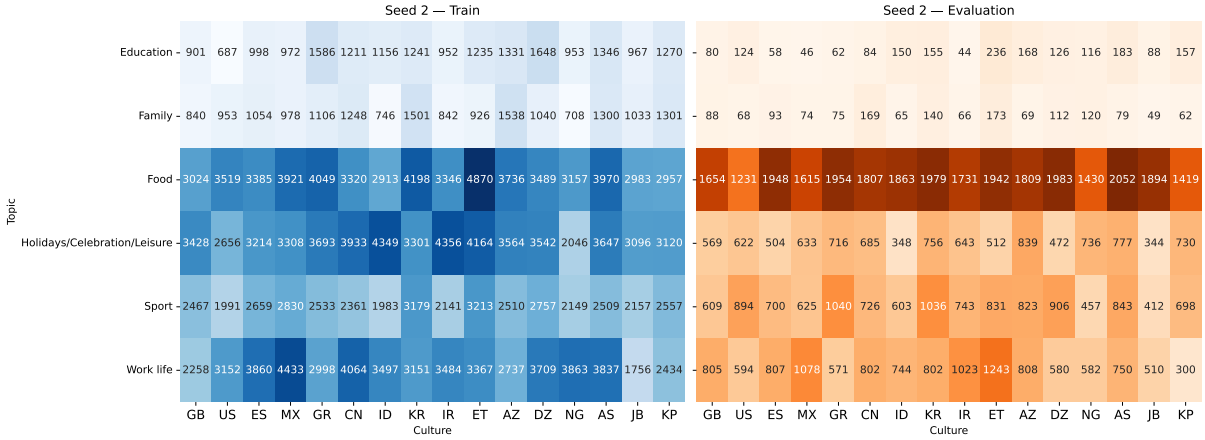


Figure 7: Topics distributions of training and evaluation questions in the data split of seed 2.

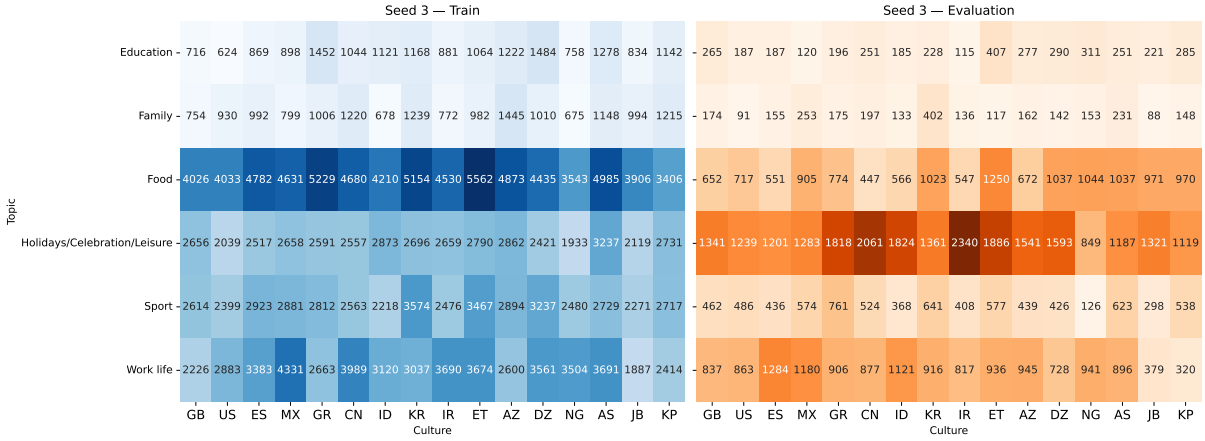


Figure 8: Topics distributions of training and evaluation questions in the data split of seed 3.

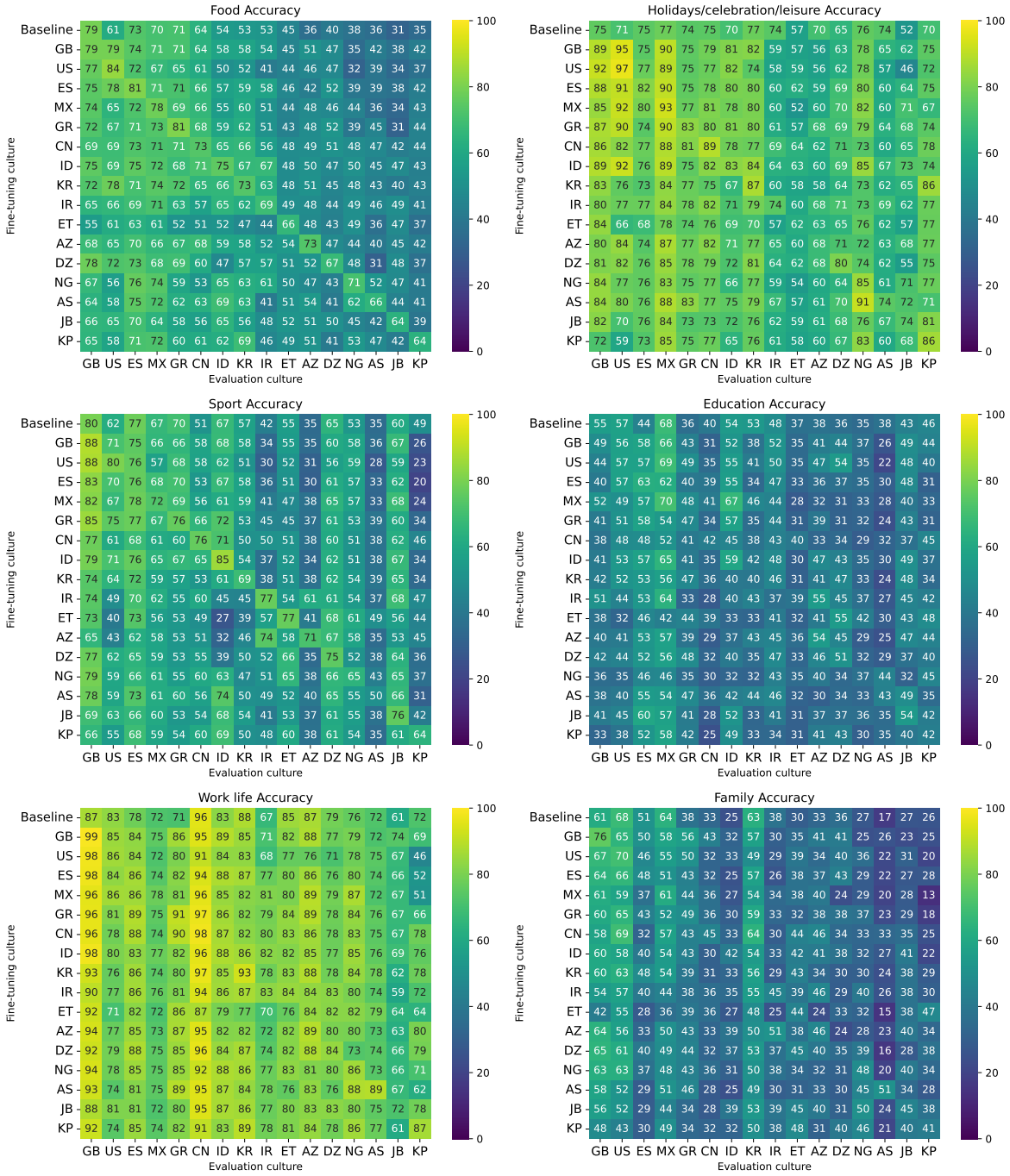
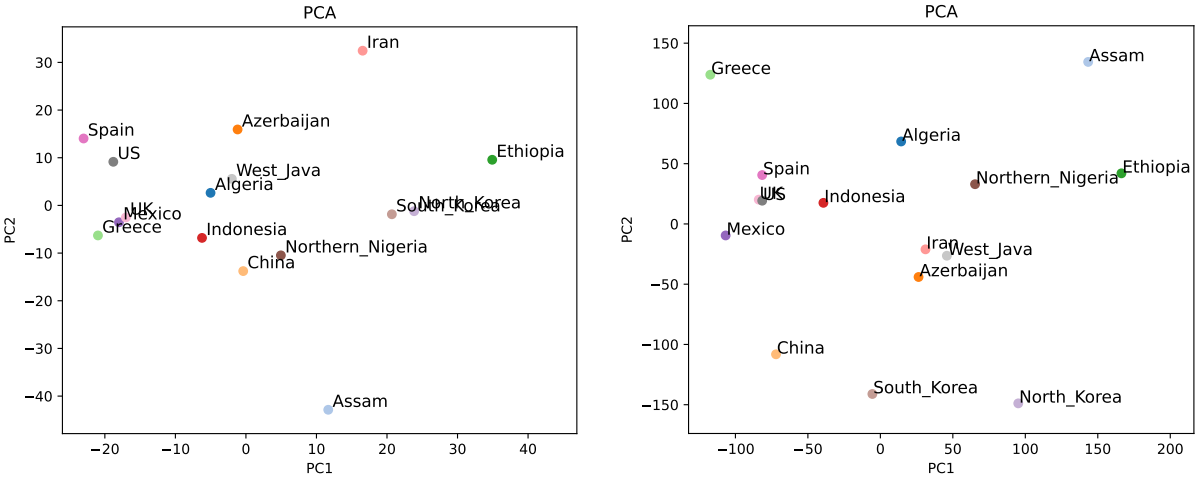


Figure 9: Prompt-tuning — Cross-cultural results across BLEND topics.



(a) Prompt-tuning

(b) Prefix-tuning

Figure 10: Visualization of the first two dimensions of PCA

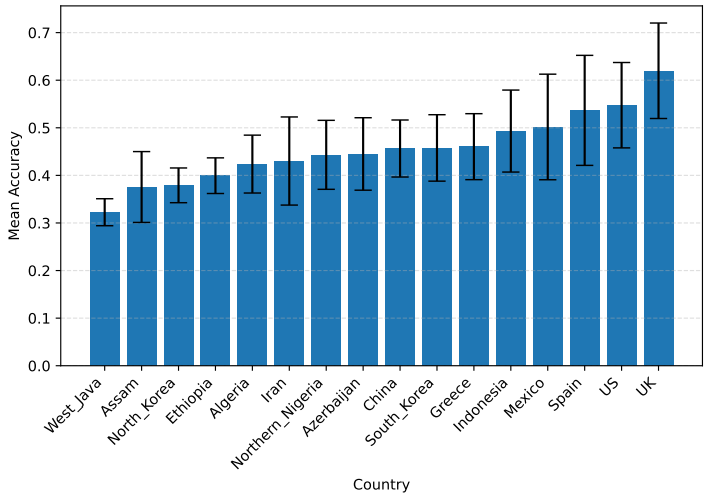


Figure 11: Ablation study — Mean accuracies and standard deviations of the models fine-tuned on the MCQ format task with prompt-tuning, evaluated on the test sets of the 16 cultures present in BLEND.