Entangled in Representations: Mechanistic Investigation of Cultural Biases in Large Language Models

Anonymous ACL submission

Abstract

The growing deployment of large language models (LLMs) across diverse cultural contexts necessitates a deeper understanding of LLMs' representations of different cultures. Prior work has focused on evaluating the cultural awareness of LLMs by only examining the text they generate. This approach overlooks the internal sources of cultural misrepresentation within the models themselves. To bridge this gap, we propose **Culturescope**, the first mechanistic interpretability-based method that probes the internal representations of different cultural knowledge in LLMs. We also introduce a cultural flattening score as a measure of the intrinsic cultural biases of the decoded knowledge from Culturescope. Additionally, we study how LLMs internalize cultural biases, which allows us to trace how cultural biases such as Western-dominance bias and cultural flattening emerge within LLMs. We find that low-resource cultures are less susceptible to cultural biases, likely due to the model's limited parametric knowledge. Our work provides a foundation for future research on mitigating cultural biases and enhancing LLMs' cultural understanding.

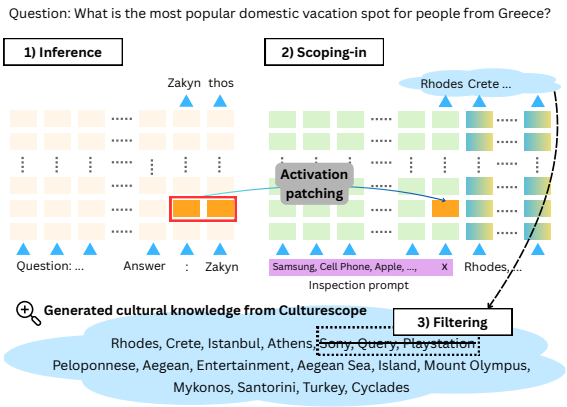


Figure 1: Given the question (about the popular domestic vacation spot in Greece), Culturescope first generates an answer to the cultural question at the **Inference** stage. Then, it reads the hidden representation from the Inference stage and elicits the cultural knowledge used for the answer ('Zakynthos') at the **Scoping-in** stage. We finalize a list of cultural knowledge after the **Filtering** stage. Culturescope unveils the internal mechanism of LLMs that cannot be revealed through the Inference stage alone.

1 Introduction

Social scientists have long observed that globalization often comes at a cost to cultural diversity. This phenomenon, known as 'Cultural Homogenization,' occurs when dominant cultures standardize and overgeneralize others, effectively erasing the distinctiveness between cultures (Pieterse, 1996; Amin, 2024). Unfortunately, this imbalance is also reflected in the pre-training data of LLMs, which is predominantly Western-centric (Santurkar et al., 2023). While LLMs do acquire some cross-cultural knowledge during training (Hershcovich et al., 2022; Arora et al., 2023), the imbalance of the sources of cultural knowledge often leads to cultural bias. This results in cultures being either marginalized through underrepresentation or 'flattened', wherein their unique traits are conflated with those of dominant cultures. (Nguyen et al., 2023a). For example, when asked about a popular leisure activity for retired men in Azerbaijan, GPT-4 responded with chess, which is a plausible answer, but one that reflects broad post-Soviet stereotypes rather than culturally specific knowledge about Azerbaijan (Myung et al., 2024).

These patterns of overgeneralization contribute to what we term 'cultural flattening': the model's tendency to conflate distinct cultural characteristics into dominant cultures. While evaluating model outputs, which corresponds to the extrinsic evaluation, is vital for detecting the presence of such overgeneralization, this approach alone cannot reveal the underlying mechanisms that produce it. It shows us that a model is biased, but not how this

bias is encoded in its parameters.

To facilitate the examination of the underlying mechanisms, we propose to study LLMs' cultural biases with mechanistic interpretability (MI) techniques. MI techniques provide us with methods that can directly examine how cultural biases discovered by the extrinsic evaluation (Santurkar et al., 2023) is internally processed within model representations, revealing where and how harmful generalizations emerge. We are the first to propose an approach for intrinsic cultural bias evaluation.

In this work, we introduce **Culturescope**, a method to probe internal representations and surface the cultural biases in play (Sec. 4.1). Figure 1 illustrates an overview of Culturescope. To examine the intrinsic cultural flattening embedded in the parameter space, we introduce a cultural flattening (CF) score, which quantifies the degree of intersection between cultural knowledge decoded by Culturescope (Sec. 4.2). We implement our framework on two cultural understanding tasks, cultural commonsense Question Answering (QA) and extractive QA, across three different LLMs.

We further challenge the model's cultural understanding by creating multiple-choice questions (MCQs) with hard negatives (Sec. 3.2.3). Culturally nuanced answers from high-resource cultures or geographically proximate countries are selected as hard negative options to simulate the cultural biases. This setup prevents LLMs from relying on surface-level elimination strategies based on the overgeneralization (Khan et al., 2025). Analyzing selected options by LLMs with the attention map method (Yuksekgonul et al., 2024) allows us to examine whether extrinsic and intrinsic cultural biases align, by revealing which options the model internally attends to (Sec. 5.1).

Our experiments reveal that LLMs internalize a Western-dominance bias and cultural flattening in their cultural knowledge space, as indicated by asymmetric CF scores between cultures (Sec. 6.1.1) and attention contribution scores (Sec. 6.1.2). This finding is further supported by extrinsic evaluations using MCQs with hard negatives, where LLMs tend to select culturally biased options when answering incorrectly (Sec. 6.2.1). Interestingly, this susceptibility to cultural bias appears weaker for low-resource cultures, likely due to their limited representation in the model's parametric knowledge. Our findings offer valuable insights for future research aimed at mitigating internalized cultural biases and developing culturally aligned LLMs.

2 Related Work

Evaluating Cultural Understanding of LLMs Previous work has proposed evaluation datasets and frameworks to assess LLMs' cultural understanding ability acquired during pre-training (Keleg and Magdy, 2023; Naous and Xu, 2025; Pawar et al., 2025). BLEnD (Myung et al., 2024) provides a multilingual commonsense QA dataset spanning 16 countries and regions, designed to uncover cross-cultural disparities in everyday knowledge. CAMeL (Naous et al., 2024) compares LLM behavior in Arabic versus Western settings across tasks like story generation, NER, and sentiment analysis, exposing systematic cultural biases in LLMs. Other multilingual benchmarks (Zhou et al., 2025; Hasan et al., 2025; Wang et al., 2024a; Cao et al., 2024) construct culturally localized evaluation datasets that span domains such as cuisine, proverbs, news, and reasoning. Across these datasets, performance gaps are consistently observed between high-resource and underrepresented languages and cultures, often linked to pretraining data imbalances that favor dominant regions (Naous and Xu, 2025).

While these efforts highlight important cross-cultural disparities, they perform an extrinsic evaluation, overlooking the underlying mechanism and cultural knowledge space embedded in LLMs. To address this gap, our paper aims to reveal how culture is embedded, entangled, or flattened within the models' inner representations.

Mechanistic Interpretability MI techniques are developed to explain the inner workings of LLMs by identifying responsible model components, such as neurons and attention heads (Meng et al., 2022; Geva et al., 2023; Yu et al., 2024). Leveraging their transparency, recent studies have employed MI techniques to investigate how specific behaviors emerge in LLMs. For instance, they have been used to uncover and manipulate components associated with social biases, enabling both diagnostic and steering interventions (Liu et al., 2024; Durmus et al., 2024; Yang et al., 2024). Despite growing interest in the cultural capabilities of LLMs, no prior work has explored cultural biases through the lens of MI. Our study fills this gap by applying MI techniques to probe the internal representation of cultural knowledge in LLMs, offering new insights into how cultural understanding is encoded and organized within the model.

3 Experimental Setup

3.1 Preliminaries

For each culture of interest y, a dataset $D = \{(q_0, C_0), ..., (q_{N-1}, C_{N-1})\}$ consists of N tuple instances that contain a question q and an option list $C = [c_0, c_1, c_2, a]$ of one gold answer a and three other options for MCQ. For MCQ, an LLM is given C and q to generate an output $O = [o_0, ..., o_{P-1}]$ consisting of P tokens. For open-ended QA, an LLM is only given q to generate the output O.

To generate an answer, an LLM converts a tokenized input text $T = [t_0, ..., t_{S-1}]$ containing S number of tokens into d-dimenstional vectors using the embedding matrix $E \in \mathbb{R}^{|\mathbb{V} \times d|}$, where $\mathbb{V}$ denotes the number of vocabularies. Then, the vectors are processed through L layers, each containing a multi-head self-attention (MHSA) layer and an MLP layer. The hidden representation x_i^l from a layer l, on a token t_i is computed by:

$$x_i^l = x_i^{l-1} + a_i^l + m_i^l \quad (1)$$

where a_i^l is an output from the MHSA layer and m_i^l from the MLP layer. The hidden representation from the last layer x_i^L is converted into a token by calculating the logits with the unembedding layer.

3.2 Datasets

3.2.1 Cultural QA Datasets

We select BLEnD (Myung et al., 2024), a cultural commonsense QA dataset, and CAMeL-2 (Naous and Xu, 2025), an extractive QA dataset featuring culturally grounded entities. BLEnD (Myung et al., 2024) is a hand-crafted multilingual benchmark designed to evaluate LLMs' everyday knowledge across diverse cultures. CAMeL-2 (Naous and Xu, 2025) is a bilingual benchmark originally constructed to evaluate LLMs' entity extraction capabilities on Arabic and English entities. We use English and Spanish questions from BLEnD and use both Arabic and English from CAMeL-2. For both datasets, we select 14 cultural groups to keep a similar culture distribution between two datasets. Dataset details can be found in Appendix A.

3.2.2 Grouping of Cultures

We categorize 14 countries from each dataset along the resource dimension and the region dimension to study how overgeneralization manifests across these dimensions. For the resource dimension, we adopt the taxonomy proposed by Joshi et al. (2020), which we simplify into three resource levels. For the region dimension, we group countries into six regions based on continents. For more details, we refer to Appendix A.1.

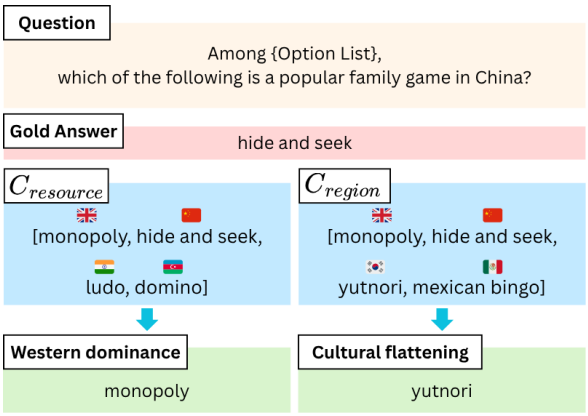


Figure 2: A cultural question about the popular family game in China from the BLEnD dataset (Myung et al., 2024). For the given question about China, if an LLM answers 'Yutnori', a popular family game in South Korea, it is caused by the effect of cultural flattening between South Korea and China. On the other hand, if the answer is 'Monopoly', the LLM is generating an answer from a high resource culture.

3.2.3 Cultural MCQ with hard negatives

Khan et al. (2025) found that if MCQs lack the adversarial depth to probe genuine cultural understanding, models can exploit surface-level elimination strategies without understanding cultural distinctions. Thus, we propose a *cultural MCQ with hard negatives* to study how overgeneralization—driven by regional or resource dominance or similarity—affects the downstream task. Since BLEnD (Myung et al., 2024) provides different answers from each culture with the same question, we create BLEnD-resource and BLEnD-region partition using culturally nuanced answers in BLEnD.

For dataset creation, we design two types of multiple choice question option lists that incorporate hard negative options: $C_{resource}$ and C_{region}, corresponding to BLEnD-resource and BLEnD-region, respectively. For $C_{resource}$, given a question q targeting culture y, we sample one culture from each of the three resource levels excluding y. We obtain these three cultures' respective gold answers when substituted into q for y, resulting in three hard negative options: c_{high}, c_{mid}, and c_{low}. For C_{region}, we sample one culture from the same geographical region as y (excluding y) and extract its corresponding answer to construct a region-based hard negative option, $c_{sameregion}$. Two additional

options ($c_{region1}$, $c_{region2}$) are randomly selected from countries in different regions. We report descriptive statistics of the number of selected countries to demonstrate the balanced sampling in Appendix A.2. All options are shuffled to avoid positional bias.

Figure 2 shows the example of $C_{resource}$ and C_{region} for the question about China's popular family game. Cultural MCQ with hard negative options allows us to examine when a model generates incorrect answers to cultural questions, whether a model's cultural confusion arises from similarity in resource level or regional proximity.

3.3 Models

Application of MI methods requires full access to model weights, hence we conduct our experiments with three recent open-sourced LLMs: Meta-Llama-3.1-8B-Instruct (`Llama-3.1`, Grattafiori et al. (2024)), aya-expanse-8b (`aya-expanse`, Dang et al. (2024)), and Qwen2.5-7B-Instruct (`Qwen2.5`, Qwen Team (2024)).

3.4 Patchscope

Patchscope (Ghandeharioun et al., 2024) utilizes an LLM to generate natural language explanations of its internal representations via two forward passes with a patching operation in between. An inspection prompt serves as a probe to extract specific knowledge encoded in the model's internal states, aligned with a predefined objective, such as next-token prediction, or attribute extraction. Utilizing an LLM itself with an inspection prompt as a probing mechanism addresses key limitations of prior methods (Hernandez et al., 2024; Geva et al., 2022; Belrose et al., 2025), which often rely on predefined probing classes or suffer from limited interpretability due to sub-word tokenization.

These limitations are particulary pronounced for cultural knowledge, which is difficult to enumerate exhaustively and often spans multiple tokens (Naous and Xu, 2025). To address these challenges, we introduce a Patchscope-based method tailored for probing the cultural knowledge space. To our knowledge, this is the first work to apply interpretability techniques for investigating cultural knowledge in LLMs.

4 Probing Cultural Knowledge within Internal Layers

Probing the cultural knowledge processed by each layer for the given input provides insights into how cultural knowledge for one culture is overlapping with different cultures within the inner layers of an LLM. To translate internal representations of LLMs to natural language that reveals the cultural knowledge space, we propose Culturescope, building upon the existing interpretability method, Patchscope (Ghandeharioun et al., 2024). Culturescope consists of three stages: inference, scoping-in, and filtering. Culturescope allows us to move beyond what is observable from model responses alone, overcoming the limitation of extrinsic evaluation.

4.1 Culturescope

Step 1. Inference An LLM first encodes a tokenized input T_i of i-th instance and generates an output O_i, which is an LLM answer to an open-ended cultural QA consisting of P number of tokens. However, activation patching during the scoping-in stage requires a single-token representation of the O_i.

Since Patchscope does not consider patching with multi-tokens (Ghandeharioun et al., 2024), we adopt Bronzini et al. (2024)'s approach, originally developed for fact-checking claims, to condense an LLM's cultural answer involving multiple tokens into a single hidden representation. Specifically, to compute the representative hidden representation x_*^l, we perform the weighted sum of hidden states as in Eq. 2 for the layer l. We set the weight w_p of each token to one if it is a noun or a verb. Other token weights are set to zero. The resulting x_*^l is then patched onto the inspection prompt during the scoping-in stage.

$$x_*^l = \sum_{p=0}^{P-1} x_p^l * w_p \tag{2}$$

Step 2. Scoping-in During this stage, we utilize activation patching to elicit the cultural knowledge encoded in x_*^l, revealing the cultural knowledge space utilized for generating O_i. To do this, the LLM performs another inference using an inspection prompt, which generates a comma-separated sequence of cultural knowledge items. The inspection prompt ends with a placeholder token 'x', where patching is done following Patchscope (Ghandeharioun et al., 2024). At layer l, we replace the hidden representation at the placeholder token position with x_*^l. The inspection prompt can be found in Appendix B. Finally, we split the generated sequence into a list of cultural knowledge $\mathcal{K}_i = [k_{i,1}, ..., k_{i,j}, ...]$.

Step 3. Filtering We empirically observe that an LLM tends to generate knowledge that is not culture-specific with our inspection prompt when the patched representation lacks the cultural knowledge. Since our method aims at eliciting any cultural knowledge available within inner representations, we devise a filtering method rather than identifying the most relevant model component.

To filter out the knowledge that is unrelated to cultural knowledge, we calculate the semantic similarity as an activation score between the input text T and the generated cultural knowledge $k_{i,j}$. We take separate encoder-based semantic search models for each language to obtain the hidden representation. In Eq. 3, we compute the representation of the input text $H_T = [h_0, ..h_s, ..h_{S-1}]$ using the final hidden states from the semantic search model. The activation score $e_{i,j}$ of $k_{i,j}$ is calculated by the cosine similarity.

$$h_t^* = \frac{1}{S}\sum_{s=0}^{S-1} h_t \tag{3}$$

We keep $k_{i,j}$ when its $e_{i,j}$ is higher than the average of all e. Details for implementation is provided in Appendix C.

4.2 Cultural Flattening Score

To quantify cultural flattening, where a culture's distinctive characteristics are misrepresented by those of a dominant culture, we introduce the Cultural Flattening score (CF score). A CF score is asymmetric and is calculated for an ordered pair of cultures, a target culture y_t and a source culture y_s.

For the CF score, we first compute a chi-square contribution for each cultural knowledge decoded by Culturescope (Sec. 4.1). This chi-square contribution quantifies how strongly a given piece of knowledge is activated for a specific culture relative to an independence assumption across all cultures, thereby serving as a measure of cultural specificity. This formulation highlights the impact of culturally distinctive knowledge rather than shared cultural knowledge arising from cultural assimilation.

For knowledge k from a culture $y \in \mathcal{Y}$, the expected count under the independence assumption, $\mathbb{E}_{k,y}$, is defined as:

$$\mathbb{E}_{k,y} = \frac{N_y \cdot \sum_y g_{k,y}}{N_{\mathcal{Y}}} \tag{4}$$

where $g_{k,y}$ is the frequency of k in the culture y, and N denotes the number of knowledge generated by y or across all cultures $\mathcal{Y}$. The chi-square contribution of cultural knowledge $\mathcal{X}_{k,y}$ is then computed as:

$$\mathcal{X}_{k,y} = \frac{(max(0, g_{k,y} - \mathbb{E}_{k,y}))^2}{\mathbb{E}_{k,y}} \tag{5}$$

We only consider cases where the residuals are positive, in order to remove knowledge that is underrepresented by the culture. The higher the $\mathcal{X}_{k,y}$, the more distincitve k is. The final CF score from the target culture y_t to the source culture y_s is:

$$F(y_t \rightarrow y_s) = \sum_{k \in \mathcal{K}} \mathcal{X}_{k,y_t} \cdot \mathbb{1}[\mathcal{X}_{k,y_s} \neq 0] \tag{6}$$

where $\mathbb{1}[\cdot]$ is the indicator function. By summing chi-square contribution scores over overlapping knowledge, the CF score captures the total amount of culturally distinctive knowledge from the target culture that is also present in the source culture. Higher values indicate that a large fraction of the target culture's distinctive concepts are represented in the source culture.

5 Tracing LLMs' Internal Mechanisms for Cultural Knowledge Usage

Leveraging MCQs with hard negatives (Sec. 3.2.3), we propose to investigate how LLMs internalize Western-dominance bias and cultural flattening via the attention map (Yuksekgonul et al., 2024). This setup facilitates an analysis of whether cultural biases are also reflected in the attention mechanism, as in extrinsic evaluation, tracing the internal mechanisms for the emergence of cultural biases.

5.1 Attention Contribution Score

To examine how LLMs internally process Western-dominance bias and cultural flattening, we analyze how attention patterns are directed toward each option c in the input T using the MCQs with hard negatives (Sec. 3.2.3). Following Yuksekgonul et al. (2024)'s work that highlights the final input token as a meaningful anchor point for attention analysis, we track attention from this final input token t_s to tokens that correspond to c in an option list C. To avoid biases from MCQ (Wang et al., 2024b), we provide options without option letters.

For a tokenized input text T that contains a culture-specific question q and a curated option list, we compute the attention contribution a_{t_c,t_s}

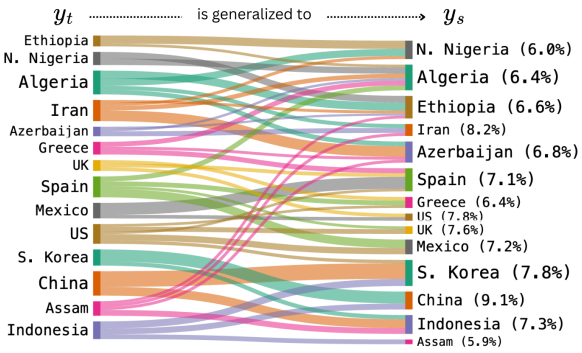


Figure 3: We present the results from the CF score on BLEnD with `Llama-3.1` in English. Percentages on the right reflect relative knowledge coverage and are defined as the ratio between the number of cultural knowledge generated for a given culture and the total number of generated concepts across all cultures.

from the final input token t_s to the token t_c corresponding to an option c, using the attention weight of the token t_c and t_s. As in Eq. 7, we obtain a layer-wise attention contribution score $a^l_{t_c,t_s}$ on layer l, where $ATTN_{t_c,t_s}$ denotes attention operation between the token t_c and t_s with a hidden representation $x^{l-1}_{t_c}$ from the previous layer.

$$a^l_{t_c,t_s} = ATTN_{t_c,t_s}(x^{l-1}_{t_c}) \tag{7}$$

The final *attention contribution score*, a_{t_c}, is computed by averaging the layer-wise scores $a^l_{t_c,t_s}$ across all layers in the LLM.

6 Experimental Results

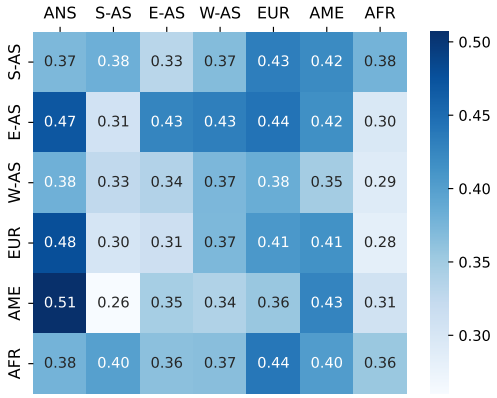
6.1 Intrinsic Cultural Bias Evaluation

6.1.1 CF Score Results

We present CF score results using a Sankey diagram, where the flow direction indicates target-to-source relations and flow width is proportional to the CF score. Figure 3 shows cultural pairs with high CF scores, where strand width reflects the extent to which culturally distinctive knowledge from a target culture y_t is generalized to a source culture y_s in the model's internal representations.

Despite their large number of generated concepts, the US and the UK exhibit relatively weak incoming connections as source countries, indicating that the CF score successfully downweights overlap driven by broadly shared cultural knowledge. The result reveals asymmetric connections, often aligned with geographic or cultural proximity, suggesting uneven generalization of culturally distinctive knowledge (China -> S. Korea, Iran -> Azerbaijan). In addition, cultures with lower concept coverage, such as Assam, Azerbaijan, and Ethiopia, consistently show narrow connections. This pattern of asymmetry is also shown in other languages. Additional results using CAMeL-2 and other models, covering English, Arabic, and Spanish, are provided in Appendix E.

(a) BLEnD-Resource



(b) BLEnD-Region

Figure 4: We present a heatmap of attention contribution scores (z-score normalized) for incorrect predictions of `Llama-3.1`. The x-axis groups correspond to the culture of the chosen option, while the y-axis groups correspond to the culture of the gold answer. For example, in Figure 4(a), the High-Low cell (0.44) indicates the average attention contribution to option tokens from the High resource group when the gold answer belongs to the Low resource group.

6.1.2 Attention Contribution Score Results

Figure 4 presents the average of attention contribution scores on option token positions, assigned by `Llama-3.1` when the model makes incorrect predictions. We separate the analysis between correct predictions and incorrect predictions, as we are particularly interested in LLMs' internal patterns when they are making biased predictions. Details for the aggregation method and results from `aya-expanse`

		BLEnD-Resource				BLEnD-Region			
		Acc	% Biased	% Others	Refusal	Acc	% Biased	% Others	Refusal
Llama-3.1	**avg.**	0.43	0.19**	0.18	0.02	0.43	0.20**	0.18	0.02
	mid	0.44	0.19	0.18	0.02	0.45	0.21**	0.16	0.02
	low	0.39	0.21	0.19	0.02	0.37	0.19	0.21	0.03
aya-expanse	**avg.**	0.38	0.18**	0.16	0.11	0.35	0.19**	0.17	0.13
	mid	0.40	0.17	0.16	0.11	0.37	0.20**	0.15	0.13
	low	0.31	0.20	0.19	0.12	0.30	0.19	0.19	0.13
Qwen2.5	**avg.**	0.44	0.20**	0.16	0.05	0.44	0.20**	0.16	0.05
	mid	0.44	0.19	0.17	0.04	0.47	0.19**	0.15	0.04
	low	0.40	0.21**	0.17	0.06	0.36	0.21**	0.18	0.06

Table 1: Model outputs result from `Llama-3.1`, `aya-expanse`, and `Qwen2.5` on the BLEnD-Resource and BLEnD-Region dataset (Sec. 3.2.3), evaluated using four metrics. Results are aggregated by question's target culture type: all averaged (avg.), mid-resource (mid), and low-resource (low). ** denotes statistically significant difference between % Biased and % Others.

and `Qwen2.5` are shown in Appendix F.

In BLEnD-Resource and BLEnD-Region, the x-axis groups correspond to the groups of the chosen option, while the y-axis groups represent the target culture of the question—resource level (High, Mid, Low) in BLEnD-Resource and region (South Asia [S-AS], East Asia [E-AS], West Asia [W-AS], Europe [EUR], America [AME], Africa [AFR]) in BLEnD-Region. From Figure 4 (a) and (b), attention contributions from the last input token to incorrect options reflect the induced biases, Western-dominance bias in BLEnD-Resource and cultural flattening in BLEnD-Region. This trend is also shown in two additional models (Figure 8). Student's t-tests indicate that the average differences in attention scores are statistically significant. Details are provided in Appendix D.2.

6.2 Extrinsic Cultural Bias Evaluation

6.2.1 Performances on MCQ with Hard Negatives

Table 1 shows the extrinsic evaluation results from `Llama-3.1`, `aya-expanse`, and `Qwen2.5` on the BLEnD-Resource and BLEnD-Region (Sec. 3.2.3). We aggregate the results by question's target culture type: all averaged (avg.), mid-resource (mid), and low-resource (low). We present the percentage of instances where the model chooses a gold answer, denoted as Accuracy (Acc). The metric labeled as "% Biased" indicates the proportion of instances where the model chooses a hard negative option, which represents a targeted bias. "% Others" represents the proportion of instances where the model chooses one of the remaining random options. Since there are two random options present in the option list, we divide the proportion of choosing random options by two for a fair comparison. We also report the proportion of instances as 'Refusal' where LLMs avoid answering.

We observe that the accuracy decreases for low-resource target questions compared to the average, which is statistically significant (Table 7). We further statistically test whether the lower accuracy in low-resource cultures stems from a preference for biased options over others. However, while the differences between % Biased and % Others are statistically significant in avg. and mid-resource cultures, the differences in low-resource cultures are not statistically significant. Details for the statistical tests are described in Appendix D.1.

7 Evaluation of Culturescope

Considering that Culturescope elicits cultural knowledge from all layers without pinpointing any specific model component, we design two validation approaches including relevance evaluation and irrelevant patching. First, for the relevance evaluation, we compare our method to three different input schemes, which are designed to enhance LLMs' cultural understanding ability. Specifically, For Culturescope, we prepend the comma-separated cultural knowledge K to the input text T and evaluate an LLM's answer O by exact-match. Table 2 shows the accuracy from different input schemes on BLEnD. Results on CAMeL-2 are in Appendix 9.

Baseline is an input without an explicit instruction and additional cultural knowledge. Cultural Prompting (Li et al., 2024; Cheng et al., 2023) is an input with an additional instruction (e.g. "The following question is about culture y") designed to guide an LLM with an explicit culture name.

Relevance evaluation			
	Llama-3.1	aya-expanse	Qwen2.5
Baseline	0.4848	0.4683	0.4626
Cultural Prompting	0.4699	0.4638	0.4664
CANDLE	0.4007	0.2692	0.3473
Culturescope	**0.5462**	**0.4928**	**0.5059**
Irrelevant patching			
	Llama-3.1	aya-expanse	Qwen2.5
English	0.325 ±0.01	0.349 ±0.1	0.344 ±0.11
Irrelevant	0.149 ±0.09	0.127 ±0.07	0.141 ±0.09

Table 2: We present the evaluation of Culturescope on BLEnD (Myung et al., 2024) in English. For the relevance evaluation results, we highlight the best performing method in bold, and the second-best with underline. For the irrelevant patching results, we report the mean along with the standard deviation.

CANDLE (Nguyen et al., 2023b) is a comprehensive cultural commonsense knowledge base which provides cultural concepts derived from assertions. The input example can be found in Appendix B.

Since topics in CANDLE and those in the datasets are not aligned, we sample 20 concepts for each culture. Due to the sampling, we perform the inference with three different random seeds and report the average accuracy. Experimental results show that inputs augmented with Culturescope perform better than those with CANDLE concepts. The accuracy with Culturescope is best on BLEnD across all models, on par on CAMeL-2 dataset with the baseline. This confirms that Culturescope reveals a valid cultural knowledge space, which is highly relevant to the given input.

For the irrelevant patching test, instead of patching representative hidden representations, we patch randomly sampled Gausian noise vectors and decode knowledge using the same inspection prompt. We report the average semantic similarity score of the generated knowledge computed during the filtering stage. Higher similarity scores for Culturescope validate that the generated cultural knowledge is not an artifact induced by the experimental setup (Table 2; Table 8). In addition, we provide the average number of generated knowledge per layer in Appendix C.

8 Discussion

Our proposed methods for evaluating intrinsic cultural biases in LLMs reveal that models internalize two cultural biases: Wester-dominance bias and cultural flattening. As shown in Figure 3, we observe asymmetric connections among geographically or culturally proximate cultures. These asymmetries indicate uneven generalization of culturally distinctive knowledge, consistent with potential cultural flattening effects. Attention-based analyses further reveal the presence of Western-dominance bias that is not captured by CF scores as the CF score explicitly downweights globally shared knowledge (Section 6.1.2; Figure 4). This highlights that the CF score and attention contribution scores capture different but complementary aspects of cultural bias in model representations.

We also observe narrow connections for cultures with low coverage, such as Assam, Azerbaijan, and Ethiopia (Section 6.1.1). Rather than indicating cultural erasure through dominance by other cultures, this pattern likely reflects limited representational coverage of culturally distinctive knowledge of LLMs. In this sense, low-resource cultures may appear less susceptible to cultural flattening. This interpretation aligns with Li et al. (2025), who report weaker memorization of symbols associated with low-resource cultures in LLMs.

Further evidence for this explanation comes from downstream behaviour: models exhibit a reduced preference for biased options involving low-resource cultures (Table 1), with the differences being statistically significant (Table 7). Taken together, these findings suggest that weaker cultural biases toward low-resource cultures do not stem from improved fairness, but rather from insufficient cultural representation. Consequently, improving model performance for low-resource cultures may require approaches that prioritize knowledge acquisition and representational coverage, rather than bias mitigation alone.

9 Conclusion

In this work, we study how cultural biases arise from the internal representations of LLMs, by investigating LLMs' internal representations. We introduce Culturescope, a method that leverages activation patching to analyse the cultural knowledge encoded within their layers. In addition, we quantify the phenomenon of 'cultural flattening', where LLMs represent less-documented cultures through the knowledge of more dominant or geographically close ones, thereby erasing cultural nuances. Our findings suggest that future work should develop a tailored approach that considers the impact of bias and resource levels to improve LLMs' cultural understanding.

Limitations

Although a growing number of benchmarks aim to evaluate cultural knowledge in LLMs, few are suitable for our evaluation setup. To meaningfully compare the probed cultural knowledge across cultures, the datasets must maintain a consistent QA format across different cultures. Thus, our methods and analyses inherit the limitations from the benchmarks, such as, lack of fine-grained set of cultures, and 1-to-1 mapping of a culture and a country. However, we emphasize that our proposed method is model- and task-agnostic, and can be applied to any dataset that meets these requirements.

Due to computational constraints, we are reporting results with 8B models, unable to conduct experiments on larger-scale models. Our definition of low-resource cultures relies on the approximation of the most-spoken language from the previous work (Joshi et al., 2020). Finally, our MCQ with hard negatives (Sec. 3.2.3) involves a degree of random sampling. While this introduces some variability, we consider it a reasonable trade-off given the prohibitive cost of exhaustively evaluating all possible negative combinations. We mitigate this by ensuring consistency across runs and focusing on aggregate trends rather than individual instances.

References

Tahir Amin. 2024. *Kashf Journal of Multidisciplinary Research*, 1(03):10–20.

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

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

Marco Bronzini, Carlo Nicolini, Bruno Lepri, Jacopo Staiano, and Andrea Passerini. 2024. Unveiling LLMs: The evolution of latent representations in a dynamic knowledge graph. In *First Conference on Language Modeling*.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, Oliver Rausch, Saffron Huang, Sam Bowman, Stuart Ritchie, Tom Henighan, and Deep Ganguli. 2024. Evaluating feature steering: A case study in mitigating social biases.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022. LM-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–21, Abu Dhabi, UAE. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: a unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Md Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural query for LLMs. In

Findings of the Association for Computational Linguistics: ACL 2025, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2151–2165.

Huihan Li, Arnav Goel, Keyu He, and Xiang Ren. 2025. Attributing culture-conditioned generations to pretraining corpora. In *The Thirteenth International Conference on Learning Representations*.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling*.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. 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 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023a. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023b. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference*.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–98.

Jan Nederveen Pieterse. 1996. Globalisation and culture: Three paradigms. *Economic and Political Weekly*, 31(23):1389–1393.

Qwen Team Qwen Team. 2024. Qwen2.5: A party of foundation models.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and

Barbara Plank. 2024b. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. 2024. Mitigating biases for instruction-following language models via bias neurons elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9061–9073, Bangkok, Thailand. Association for Computational Linguistics.

Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. 2024. Revealing the parametric knowledge of language models: A unified framework for attribution methods. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8173–8186, Bangkok, Thailand. Association for Computational Linguistics.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *The Twelfth International Conference on Learning Representations*.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025. Does mapo tofu contain coffee? probing LLMs for food-related cultural knowledge. 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 9840–9867, Albuquerque, New Mexico. Association for Computational Linguistics.

Dimension	Groups	Countries
Resource	High	Algeria, China, Iran, Mexico, Spain, UK, US
	Mid	Greece, Indonesia, South Korea
	Low	Assam, Azerbaijan, Ethiopia, Northern Nigeria
Region	South Asia	Assam, Indonesia
	East Asia	China, South Korea
	West Asia	Azerbaijan, Iran
	Europe	Greece, Spain, UK
	North America	Mexico, US
	Africa	Algeria, Ethiopia, Northern Nigeria

Table 3: Culture groups and the list of cultures belong to each group

BLEnD-Resource		BLEnD-Region	
High	2548	Europe	789
		North America	520
Mid	1425	East Asia	588
		West Asia	516
Low	1077	South Asia	592
		Africa	771
Sum	**5050**	**Sum**	**3776**

Table 4: Number of questions per group in Cultural MCQ with hard negatives

Option type	Mean ± s.d.	95% CI
BLEnD-Region		
Gold answer	269.71 ± 16.49	[260.2, 279.23]
Hard negative	269.71 ± 16.3	[260.3, 279.12]
Others	539.43 ± 38.34	[517.3, 561.56]
BLEnD-Resource		
Gold answer	360.71 ± 11.4	[354.13, 367.30]
High	721.43 ± 31.55	[692.25, 750.61]
Mid	1683.33 ± 52.5	[1553.9, 1812.78]
Low	1262.5 ± 85.57	[1126.49, 1398.51]

Table 5: Descriptive statistics (mean, standard deviation (s.d.), and 95% Confidence Interval (CI)) of the number of selected culture groups for each option type in BLEnD-Resource and BLEnD-Region.

A Dataset Details

In our experiments, we utilize BLEnD (Myung et al., 2024) and CAMeL-2 (Naous and Xu, 2025). BLEnD comprises 500 short-answer question-answer pairs for each culture, where the answers vary depending on the cultural or regional context. To reduce computational cost, we exclude North Korea and West Java, resulting in a final selection of 14 cultural groups from BLEnD.

With CAMeL-2, an LLM is asked to extract an entity from a context collected from Arabic X/Twitter data according to the specified entity type in the input. We take the English partition and reduce the dataset to 14 countries to keep a similar culture distribution to the BLEnD dataset. Dataset details, including domains and countries covered by the datasets, can be found in Appendix A.

We provide their brief data statistics and characteristics in Table 6.

A.1 Culture Groups

As mentioned in Sec. 3.2, our work conducts experiments that focus on 14 countries classified in two dimensions. Specifically, with the taxonomy from Joshi et al. (2020) which classifies languages into six levels (0: very low-resource to 5: very high-resource), we simplify this into three groups: High (Level 5), Mid (Levels 3-4), and Low (Levels 0-2). We assign each culture a language resource level based on its most widely spoken language provided by Wikipedia.

For the region dimension, we group countries into six regions based on continents. We split Asia into three subregions, which leaves us six regional groups: North America, Europe, Africa, West Asia, South Asia, and East Asia. We compare three groups based on the level of language resource (Joshi et al., 2020) and six groups based on the continental region. Table 3 shows the culture entities that correspond to each group.

A.2 Cultural MCQ with hard negatives

The creation of BLEnD-Resource and BLEnD-Region involve random selection among possible options (Sec. 3.2.3). Table 4 presents descriptive statistics including mean, standard deviation, and 95% confidence interval of the number of selected culture groups for each option type. This demonstrates that options are sampled to achieve balanced representation of cultural groups. An option type of gold answer refers to a culture group from which the gold answer option originates. An option type of hard negative refers to a culture group from which an option selected as a hard negative originates. High, mid, and low denote the number of selected countries from three resource groups.

Dataset	Task	Number of Questions	Domain	List of Countries
BLEnD Myung et al. (2024)	Cultural Commonsense QA	5,726	Education, Food, Holidays/Celebration/Leisure, Sport, Work life, Family	Africa: Algeria, Ethiopia, Northern Nigeria Europe: Spain, United Kingdom, Greece North America: United States, Mexico East Asia: China, South Korea South Asia: Indonesia, Assam West Asia: Iran, Azerbaijan
CAMeL-2 Naous and Xu (2025)	Extractive QA	1,862	Locations, Beverage, Food, Sports	Africa: Morocco, Algeria Europe: Spain, United Kingdom, Greece North America: United States, Mexico East Asia: China, Japan South Asia: Indonesia, India West Asia: Iran, Syria, Egypt

Table 6: Details for the datasets

	BLEnD-Resource									BLEnD-Region								
	Hypothesis 1			Hypothesis 2			Hypothesis 3			Hypothesis 1			Hypothesis 2			Hypothesis 3		
`Llama-3.1`		low	mid	avg.	**0.008**			low			low	mid	avg.	**0.0049**				low
	avg.	**0.0034**	0.6328	mid	0.2256	avg.	0.1066			avg.	**0.0004**	0.931	mid	**0.0016**		avg.	0.7306	
	mid	**0.0107**	-	low	0.076	mid	0.1008			mid	**0.0001**	-	low	0.8067		mid	0.8815	
`aya-expanse`		low	mid	avg.	**0.009**			low			low	mid	avg.	**0.0124**				low
	avg.	**0.0**	0.936	mid	0.1477	avg.	0.0946			avg.	**0.0007**	0.8161	mid	**0.0086**		avg.	0.4606	
	mid	0.0	-	low	0.2307	mid	0.0548			mid	**0.0007**	-	low	0.5762		mid	0.6784	
`Qwen2.5`		low	mid	avg.	**0.0**			low			low	mid	avg.	**0.0**				low
	avg.	**0.026**	0.461	mid	0.22	avg.	0.3283			avg.	**0.0**	0.9722	mid	**0.0122**		avg.	0.1162	
	mid	0.0796	-	low	**0.0215**	mid	0.1636			mid	**0.0**	-	low	**0.0307**		mid	0.1076	

Table 7: P-values from two-proportions Z-tests with three different hypotheses. Bolded texts mean that the test supports the hypothesis.

		BLEnD	CAMeL-2
`Llama-3.1`	English	0.325 ± 0.01	0.239 ± 0.07
	Spanish	0.277 ± 0.07	-
	Irrelevant	0.149 ± 0.09	0.155 ± 0.08
`aya-expanse`	English	0.349 ± 0.1	0.242 ± 0.07
	Irrelevant	0.127 ± 0.07	0.141 ± 0.06
`Qwen-2.5`	English	0.344 ± 0.11	0.227 ± 0.07
	Arabic	-	0.173 ± 0.07
	Irrelevant	0.141 ± 0.09	0.148 ± 0.07

Table 8: Average and standard deviations of the similarity scores from the filtering stage of the Culturescope. We round the average off to two decimal places.

CAMeL-2			
Relevance Evaluation			
	`Llama-3.1`	`aya-expanse`	`Qwen2.5`
Baseline	**0.7126**	0.7019	<u>0.6675</u>
Cultural Prompting	<u>0.6799</u>	**0.7282**	**0.7148**
CANDLE	0.5640	0.6955	0.6473
CultureScope	0.6519	<u>0.7169</u>	0.6659

Table 9: We present the relevance evaluation of Culturescope on CAMeL-2 in English. We highlight the best performing method in bold, and the second-best with underline.

B Example of Prompts

In Figure 5, we present the prompt templates we use for each method to obtain Table 2.

C Details for Culturescope

During the filtering step in the Culturescope, we use three different semantic search models for each language. For English, we use a DeBERTa based model, trained for semantic search[1]. For Arabic, we use the bert-base model finetuned for semantic search in Arabic[2]. Lastly, for Spanish, we use the bert-base model finetuned for semantic search in Spanish[3].

In Table 8, we show the threshold values selected for each dataset and language for the filtering stage. The values are selected by the average of cosine similarity score across all the generated concepts.

Figure 6 describes the average number of generated knowledge per layer. This shows that Culturescope successfully generates relevant cultural knowledge from layers, especially from early layers.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] https://huggingface.co/akhooli/Arabic-SBERT-100K

[3] https://huggingface.co/hiiamsid/sentence_similarity_spanish_es

	low vs high	low vs mid
Llama-3.1	**1.11e-09**	0.159
aya-expense	**8.55e-10**	**9.69e-13**
Qwen2.5	**1.75e-07**	**5.54e-03**

Table 10: p-value of Student's t-test for each model on low-resource target samples from BLEnD-Resource, based on attention scores. Comparisons were conducted between low and mid, low and high, all from $C_{resource}$, which are all hard negative options.

D Statistical Significance Tests

D.1 Testing on MCQ Results

To evaluate the statistical significance of our findings in Sec. 6.2.1, we set three alternative hypotheses and test with a two-proportions Z-test [4]. Table 7 presents p-values from the list of Z-tests. avg., mid, and low follow the definition from Sec. 6.2.1. The Z-test supports the alternative hypothesis when the p-value is lower than 0.05. Hypothesis 1 is that the accuracy from the avg. and the mid is higher than the accuracy from the low. The p-values in all models on both datasets support this hypothesis. Hypothesis 2 is that in each level, when an LLM predicts incorrectly, it prefers the biased option over the other options. This hypothesis is supported in avg. and mid groups, which highlights our finding on less susceptibility of low-resource cultures with cultural biases. Lastly, hypothesis 3 is that the model predicts biased options more in the low group than in the avg. and mid groups. This hypothesis is rejected in all tests, confirming the statistical significance of lower cultural bias in the low group.

D.2 Testing on the Results for Attention Contribution Score

As discussed in Sec. 6.1.2, our analyses suggest that high-resource bias and Western-dominance bias are strongly internalized within LLM representations, as visualized in the attention heatmaps in Figure 4 and Figure 8. To further validate these observations, we conduct Student's t-tests[5] on the attention scores of BLEnD-Resource and BLEnD-Region.

To examine bias patterns, we conduct p-value analyses on attention scores using Student's t-tests. For high-resource bias, low-resource targets (c_{low}) are compared against mid- and high-resource (c_{mid}, c_{high}) samples within $C_{resource}$, with most p-values less than 0.05, indicating statistically significant differences (Table 10). For Western-dominance bias, for non-Western regions (South Asia, East Asia, West Asia) targeted samples, non-Western regions ($c_{sameregion}$) are compared against Western regions (Europe, America) ($c_{region1}$, $c_{region2}$) within C_{region}. While fewer comparisons yield significance, several regions with large mean differences show p-values less than 0.05 (Table 11), suggesting Western-dominance bias is present but less uniformly across regions.

In summary, the results confirm that high-resource bias is consistently strong in BLEnD-Resource, while Western-dominance bias is also evident in BLEnD-Region, though less uniformly across all regions.

E CF Score Results

In Figure 7, we show the CF score results with Llama-3.1, aya-expanse, Qwen2.5 on BLEnD and CAMeL-2. As mentioned in Sec. 6.1.1, we exclude the countries with CF scores lower than the average CF score across all countries.

F Attention Contributions

In cases where an option consists of multiple tokens, we follow the approach of Yuksekgonul et al. (2024), taking the maximum attention contribution score among the component tokens. When the attention scores are averaged by samples to examine its general patterns, simply averaging can be sensitive to extreme values or samples in which attention scores are globally biased toward higher or lower magnitudes. This variability can potentially reduce the generalizability of the results. To address this, we applied normalization (z-score normalization) per sample to the attention contribution scores, such that the scores within each sample have a mean of 0 and a standard deviation of 1.

[4] https://www.statsmodels.org/dev/generated/statsmodels.stats.proportion.proportions_ztest.html

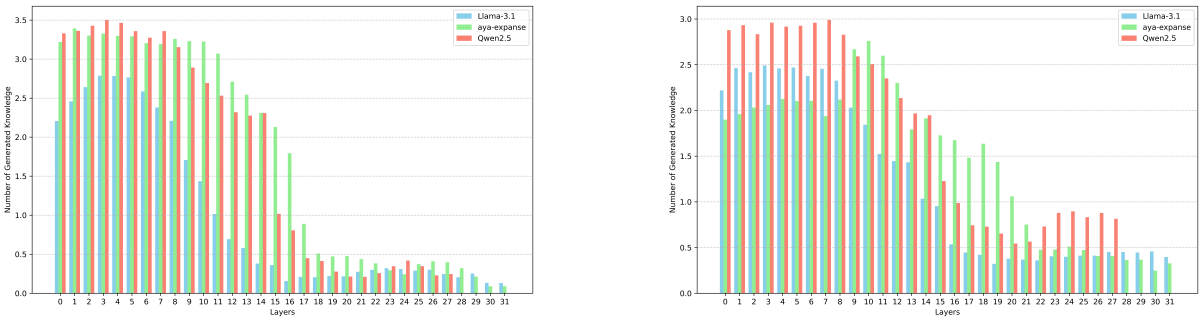
[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

Target	Llama-3.1		aya-expanse		Qwen2.5	
	Target vs EUR	Target vs AME	Target vs EUR	Target vs AME	Target vs EUR	Target vs AME
S-AS	0.0570	0.199	0.614	0.596	**0.00109**	**0.0413**
E-AS	0.864	0.403	0.0851	0.787	0.799	0.609
W-AS	0.827	0.476	0.787	0.829	0.329	0.475
EUR	**4.22e-03**	0.216	**2.26e-06**	**1.89e-04**	0.160	0.104

Table 11: p-value of Student's t-test for each model on non-Western targeted samples from BLEnD-Region, based on attention scores. Comparisons were made between each sample's target region and America, and between each sample's target region and Europe, all within C_{region}, which represents hard negative options.

1. Inspection prompts
Generate associated words, Syria, Oman, Jordan, Qatar, West Asia, Turkey, Israel, Lebanon, ..., Leonardo DiCaprio, Tom Cruise, Kate Winslet, Brad Pitt, Actor, ..., Samsung, Cell Phone, TV, Apple, Nokia, South Korea, Electronics, ..., x
2.Open-ended QA Prompts
2-1. Baseline
BLEnD
Answer the question.\n\n Question: {question} \n\n Provide your answer as "Answer: [Answer]"
CAMeL-2
Extract the {entity type} mentioned in the following text: \n\n Text: {text} \n\n Reply only with the name of the {entity type} mentioned
2-2. Cultural Prompting
BLEnD
You are given a question about {culture}. Answer the question.\n\n Question: {question} \n\n Provide your answer as "Answer: [Answer]"
CAMeL-2
You are given a question about {culture}. Extract the {entity type} mentioned in the following text: \n\n Text: {text} \n\n Reply only with the name of the {entity type} mentioned
2-3. CANDLE & Culturescope
BLEnD
You are given a question about {culture}. Answer the question, you can use list of concepts if it's relevant. \n\n Concepts: {cultural knowledge from the methods} \n\n Question: {question} \n\n Provide your answer as "Answer: [Answer]"
CAMeL-2
You are given a question about {culture}. You can use the hints if they are relevant \n\n Hints: {cultural knowledge from the methods} \n\n Extract the {entity type} mentioned in the following text: \n\n Text: {text} \n\n Reply only with the name of the {entity type} mentioned

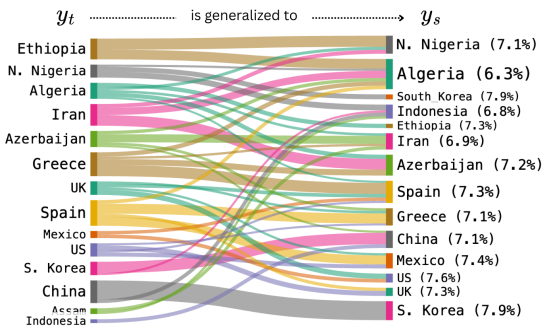
Figure 5: Prompt templates used for open-ended QA evaluations.



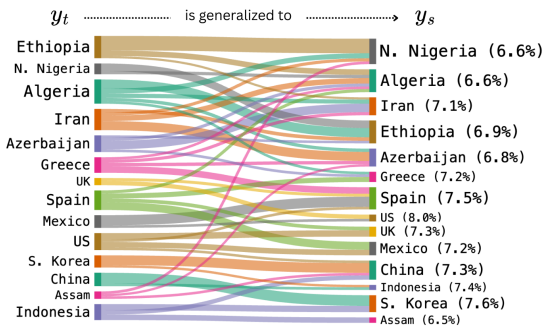
(a) BLEnD, English

(b) CAMeL-2, English

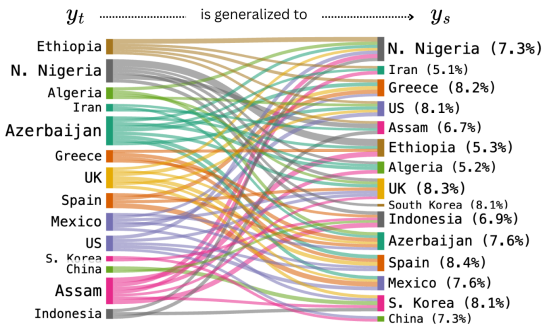
Figure 6: Average number of generated knowledge per layer.



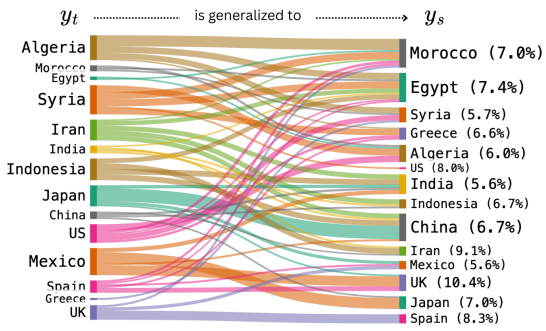
(a) BLEnD, English, aya-expanse



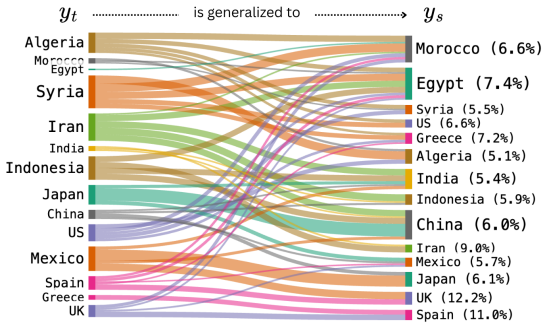
(b) BLEnD, English, Qwen2.5



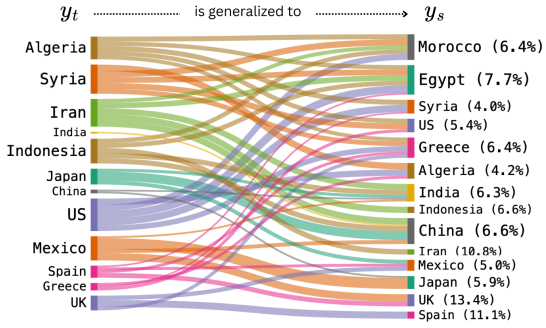
(c) BLEnD, Spanish, Llama-3.1



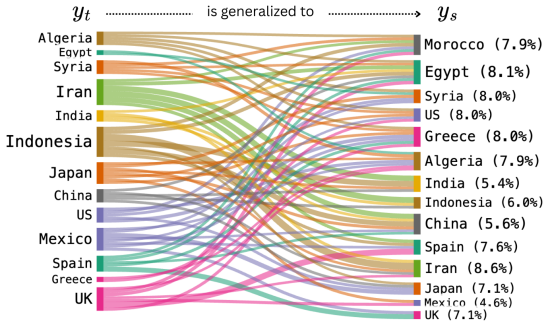
(d) CAMeL-2, English, Llama-3.1



(e) CAMeL-2, English, aya-expanse

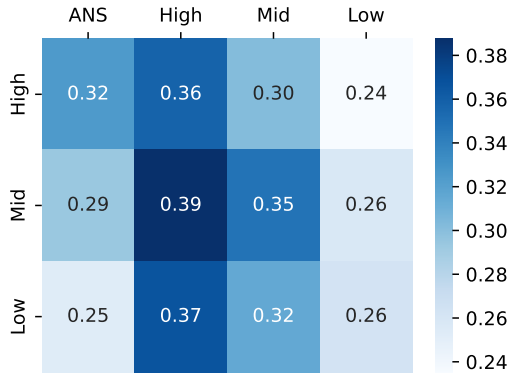


(f) CAMeL-2, English, Qwen2.5

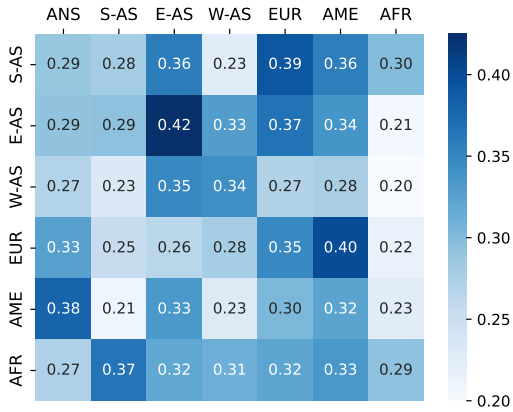


(g) CAMeL-2, Arabic, Qwen2.5

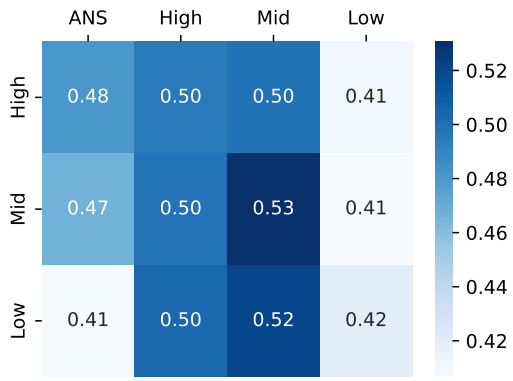
Figure 7: CF Score Results



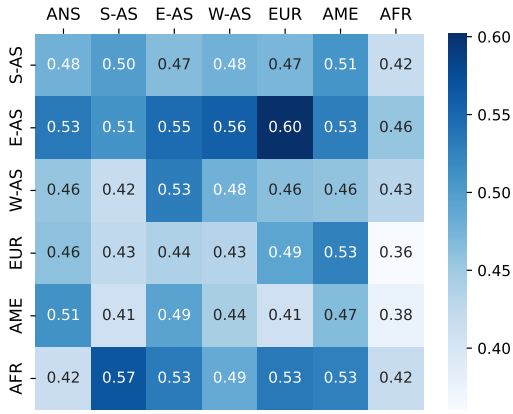
(a) BLEnD-Resource - Qwen2.5-7B-Instruct



(b) BLEnD-Region - Qwen2.5-7B-Instruct



(c) BLEnD-Resource - Aya Expanse 8B



(d) BLEnD-Region - Aya Expanse 8B

Figure 8: Heatmap visualization of average attention contribution scores (z-score normalized) on incorrect predictions.