Mechanistic Investigation of Cultural Knowledge 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' hidden representations of different cultures. Prior work has evaluated cultural awareness of LLMs through model outputs by analyzing the text they generate. This approach overlooks potential misrepresentations of cultures within the model parameters. To bridge this gap, we propose **Culturescope**, the first mechanistic interpretability-based method that probes the internal representations of LLMs for cultural knowledge. Through Culturescope, we uncover imbalanced overlaps across cultures within model parameters. We further correlate this internal imbalance with the model performance on a dataset where the cultural confusion is intentionally increased. Experimental results demonstrate that model layers do not process cultural knowledge in the same way, with greater cultural imbalance in deeper layers. Our findings provide insights into how internal representational dynamics shape the utilization of cultural knowledge in LLMs, which can inform future methods to improve the cultural alignment of LLMs.

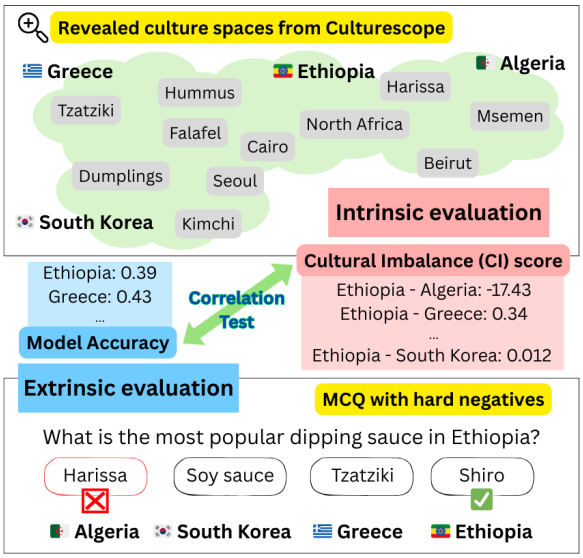


Figure 1: We propose a diagnostic framework that enables both intrinsic and extrinsic evaluation of LLMs' cultural understanding. Culturescope first reveals culture spaces for each culture by generating cultural concepts from internal representations, and a cultural imbalance (CI) score measures the imbalanced overlap between two cultures. We further correlate the CI score and model accuracy.

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 cultural distinctiveness (Pieterse, 1996; Lieber and Weisberg, 2003). A similar phenomenon can emerge in large language models (LLMs), whose pre-training data is heavily skewed toward Western sources (Santurkar et al., 2023). Although LLMs acquire broad cross-cultural knowledge during training (Hershcovich et al., 2022; Arora et al., 2023), this imbalance can lead to cultural misattributions, where culturally specific traits are incorrectly associated with related or dominant cultures (Nguyen et al., 2023). 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 reflects broad post-Soviet stereotypes rather than backgammon, which is more culturally specific (Myung et al., 2024).

Prior work has primarily studied cultural awareness through extrinsic evaluations, measuring model outputs on cultural understanding tasks (AlKhamissi et al., 2024; Bhatt and Diaz, 2024). While these approaches can reveal when models produce culturally inaccurate responses, they provide limited insight into why such errors occur. In particular, they do not explain how cultural knowledge is internally represented or how misattribu-

tions emerge within model representations. To investigate these mechanisms, we turn to mechanistic interpretability (MI), a line of work that studies how information is represented and processed within LLMs. MI techniques enable us to analyze relationships among internal representations across cultures, thereby examining how different cultures are encoded and related within a model's cultural knowledge space.

In this work, we introduce a diagnostic framework that evaluates LLMs' cultural understanding with both intrinsic and extrinsic measures. Figure 1 illustrates how our framework investigates LLMs' cultural knowledge in both intrinsic and extrinsic ways. We begin examining the model's internal representations using Culturescope, a method inspired by Patchscope (Ghandeharioun et al., 2024) but designed specifically to probe culturally associated concepts (e.g., foods, activities, or traditions) within model representations (§4.1). *By analyzing how these concepts overlap across cultures, we observe that certain cultures exhibit lower association with shared concepts than the corresponding majority cultures, a pattern similar to cultural flattening observed in the real world (§5.2).*

In order to compare cultural knowledge in model representations with downstream behaviour, we construct a multiple-choice question (MCQ) dataset containing culturally grounded questions (§6.1). Each question includes a hard negative option that targets misattributions potentially arising from overlapping culture spaces. These distracting options require models to distinguish subtle cultural differences. We further correlate model performance on the proposed dataset, measuring how the cultural imbalance observed in intrinsic evaluation informs model performance.

Our experimental results reveal that cultural knowledge is processed differently across layers, with higher CI scores in deeper layers (§4.2). We further show that cultural imbalance in model internal representations correlates with model performance on the proposed culturally challenging MCQ dataset (§6.2). *These findings suggest that cultural imbalance in model representations is associated with downstream performance differences and may emerge through layer-wise processing of cultural concepts.* Overall, our findings provide insights into how internal representational dynamics shape the utilization of cultural knowledge in LLMs, which may inform future methods to improve the cultural alignment of 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 cultures 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., 2024; 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 pre-training 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 cultural knowledge is embedded in models' internal 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., 2024a). 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 knowledge through the lens of MI. Our study fills this gap by applying MI techniques to probe the internal representations of cultural knowledge in LLMs, offering new insights into how these representations shape models' cultural understanding.

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 guides the second forward pass by showing a few-shot demonstration of an objective. For example, Ghandeharioun et al. (2024) utilize '$subject_1$: $description_1$, ..., $subject_k$: $description_k$, x' as an inspection prompt to decode an entity description from a hidden representation. Utilizing an LLM itself as a probing mechanism addresses key limitations of prior methods (Hernandez et al., 2024; Geva et al., 2022; Belrose et al., 2025), namely relying on predefined probing classes or limited interpretability due to sub-word tokenization.

These limitations are particularly pronounced for cultural knowledge, because it is difficult to enumerate all possible knowledge 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.

3 Experimental Setup

We examine models' internal representations of cultural knowledge using two cultural QA tasks.

3.1 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 is a hand-crafted multilingual benchmark designed to evaluate LLMs' everyday knowledge across diverse cultures. CAMeL-2 is a bilingual benchmark originally constructed to evaluate LLMs' entity extraction capabilities on Arabic and English entities. We use English questions from both datasets. To keep a similar culture distribution between the two datasets, we select 14 cultures to study. Dataset details can be found in Appendix A.

3.2 Models

MI methods allow us to examine internal representations when full model access is available, hence we use three popular open-weight 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)).

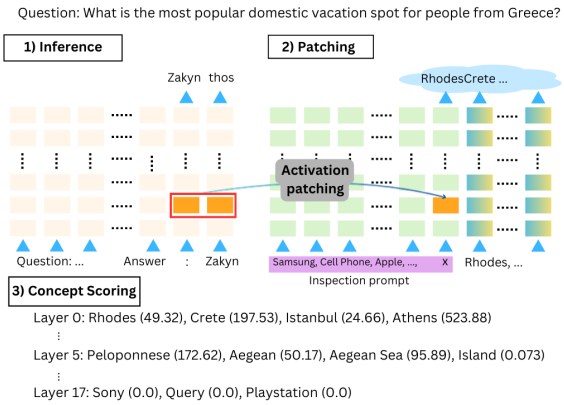


Figure 2: Given the question (about the popular domestic vacation spot in Greece), Culturescope first generates an answer to the cultural question at the **Inference** stage. Then, it reads the hidden representation from the Inference stage and elicits the cultural knowledge used for the answer ('Zakynthos') at the **Patching** stage. The **Concept Scoring** computes an association score for each concept, highlighting those most strongly associated with Greece. Culturescope reveals the internal mechanisms of LLMs that cannot be revealed by the Inference stage alone.

4 Probing Cultural Concepts within Internal Layers

Translating LLMs' internal representations into natural language can reveal knowledge stored in the cultural concept space. To this end, we propose Culturescope, which consists of three stages: inference, patching, and concept scoring. Unlike Patchscope (§2), Culturescope supports multi-token patching and includes a dedicated concept scoring stage for assessing cultural associations.

4.1 Culturescope

Step 1. Inference An LLM first encodes a tokenized cultural question q and generates an output $Y = [y_1, ..., y_P]$, which is an LLM answer consisting of P tokens. Each q is a question specific to a culture c. Since Patchscope only considers a single token for patching (Ghandeharioun et al., 2024), we calculate the weighted sum of hidden states, which is an output from a layer l as in Eq. 1.

$$\boldsymbol{x^l} = \sum_{p=1}^{P} x_p^l * w_p \tag{1}$$

Following Bronzini et al. (2024)'s approach, originally developed for condensing hidden states

Model	Culture	Top-10 concepts
Llama-3.1	South Korea	Busan, Japchae, Seoul, Kpop, Kimchi, Daejeon, Bulgogi, Incheon, Daegu, Bibimbap
	Northern Nigeria	Yoruba, Hausa, Igbo, Ghana, Nigeria, Plantain, Jollof Rice, Jollof, Senegal, Benin
	Mexico	Guacamole, Nachos, Burrito, Quesadilla, Belize, Cancun, Taco, Guatemala, Burritos, Salsa
aya-expanse	South Korea	Kimchi, Korean, Kim, Seoul, South, North Korea, Bibimbap, Kim Jongil, Korea, Kim Jongun
	Northern Nigeria	Nigeria, Ghana, Lagos, Abuja, Port Harcourt, South Africa, Senegal, Cameroon, Kano, West Africa
	Mexico	Mexican, Mexico, Sánchez, Colombia, Peru, Salsa, Taco, Tortilla, Chile, Burrito
Qwen2.5	South Korea	Busan, Incheon, Daegu, Gwangju, South Korea, Daejeon, Kimchi, Bibimbap, Korean Wave, Seoul
	Northern Nigeria	Kuli, Nigeria, Ghana, Kula, Benin, Lagos, Jollof Rice, Kano, Fufu, Cameroon
	Mexico	Burrito, Mexican Food, Quesadilla, Tacos, Texmex, Guacamole, Mezcal, Enchilada, Margarita, Enchiladas
Irrelevant patching	South Korea	Vital, Pronounced, Huge, Delete, Signed By, Wendu, Follow, , Perceived, Klip,
	Northern Nigeria	Someone, Unchanged, Affected, Aka, Seen, Yo, Everything, Han, Foces, Gazetteer
	Mexico	Feeds, M, Genius, , Swingear, Běžet, Vang, Ug, Echo, Alone

Table 1: We present the top-10 concepts per culture generated by Culturescope (Sec. 4.1) and irrelevant patching (from `Llama-3.1`) for BLEnD.

for fact-checking claims, 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.

Step 2. Patching At this stage, we use activation patching to generate a comma-separated sequence of cultural concepts encoded in $\boldsymbol{x^l}$, thereby revealing those related to Y. To do this, the same LLM performs another inference step using an inspection prompt, which includes a few-shot demonstration of listing relevant concepts. The inspection prompt is provided in Appendix B. Following Patchscope (Ghandeharioun et al., 2024), the prompt ends with a placeholder token 'x', where patching is done. At layer l, we replace the hidden representation at the placeholder token position with $\boldsymbol{x^l}$. Finally, we split the generated sequence into a list of cultural concepts. The number of concepts J may vary depending on the layer and input, as we do not generate a fixed number of concepts. We aggregate generated concepts across questions for each layer and a culture, resulting in $\mathcal{K}_c^l = [k_1^l, ..., k_j^l, ..., k_J^l]$ from layer l for culture c. A cultural concept k_j^l is a string of potentially multiple tokens. This allows us to count the frequency of a concept for a layer.

Step 3. Concept Scoring To quantify the cultural importance of each concept, we design a scoring function inspired by a chi-square contribution. The scoring function $s(k_j^l, c)$ computes an association score for each concept k_j^l for a culture c at layer l. This scoring function assigns higher scores to concepts that appear more uniquely and frequently within a specific culture while downweighing concepts that are uniformly shared across multiple cultures. As a result, culturally distinctive concepts receive higher association scores. The association scores are computed independently for each layer.

For a cultural concept k_j^l generated with a question targeting a culture c, the expected count under the independent assumption, $\mathbb{E}_{k_j^l,c}$, is defined as:

$$\mathbb{E}_{k_j^l,c} = \frac{N_c^l \cdot \sum_{c' \in \mathcal{C}} freq(c', k_j^l)}{\sum_{c' \in \mathcal{C}} N_{c'}^l} \quad (2)$$

where $freq(c, k_j^l)$ is the frequency of k_j^l in the culture c on the layer l, and $\mathcal{C}$ denotes the set of all cultures in a dataset. N_c^l denotes the number of concepts generated by c at l. The association score $s(k_j^l, c)$ of the k_j^l is then computed as:

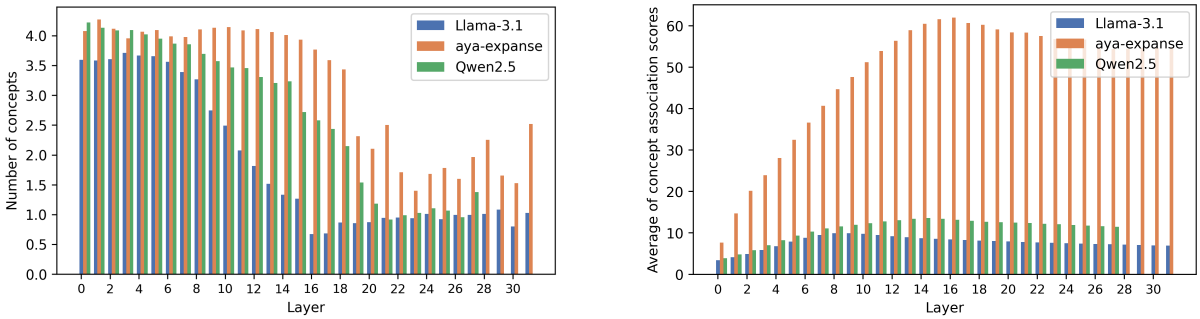
$$s(k_j^l, c) = \frac{(max(0, freq(c, k_j^l) - \mathbb{E}_{k_j^l,c}))^2}{\mathbb{E}_{k_j^l,c}} \quad (3)$$

We set the association score to zero when $freq(c, k_j^l) - \mathbb{E}_{k_j^l,c}$ is negative, to remove knowledge that is underrepresented in the culture. The higher the $s(k_j^l, c)$, the more distinct k_j^l is.

4.2 Evaluation of Culturescope

Generated Concepts from Culturescope Table 1 shows the top-10 concepts for three cultures, sorted by association scores with BLEnD. The sorting of concepts is done by averaging the association score for each concept across layers. Additional examples from other cultures on BLEnD and CAMeL-2 are in Appendix B.2.

To examine how concept generation varies across layers, we compute two statistics for each layer: (1) the average number of concepts generated per question, and (2) the average association score of the generated concepts. Figure 3a



(a) Averaged number of concepts generated per layer

(b) Averaged association scores per layer

Figure 3: We show how Culturescope outputs differs across layers w.r.t. the number of generated concepts per question (Figure 3a) and the association scores for generated concepts (Figure 3b), for BLEnD.

shows that the number of concepts generated by Culturescope decreases in deeper layers. In contrast, Figure 3b shows that the average association scores of generated concepts tend to increase or remain stable as layers become deeper. *These results suggest that earlier layers produce larger but broader cultural concepts, while later layers produce fewer but more culturally distinctive concepts.* We note that `aya-expanse` exhibits higher association scores overall, which reflects its tendency to generate fewer unique concepts and repeat them more frequently for one culture compared to the other models (Appendix B.3). The analysis on CAMeL-2 shows the same trend (Appendix B.2, Figure 8). Throughout the paper, we consider all layers to observe how they vary in intrinsic and extrinsic measures introduced later.

Irrelevant Patching To test the robustness of Culturescope, we compare concepts generated by Culturescope with those generated by patching irrelevant vectors. As an irrelevant vector representation, we create randomly sampled Gaussian noise vectors and decode concepts using the same inspection prompt. Concepts generated from irrelevant patching are shown in Table 1. To evaluate generated concepts, we calculate the cosine similarity between concepts and the gold answer to the input question that we obtained representation from using a sentence embedding model.[1] We report the average semantic similarity scores across questions that are used during the inference. We observe higher similarity scores for Culturescope, validating that the generated cultural concepts are not an artifact induced by the experimental setup (Appendix B.4).

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

5 Analyzing Internal Cultural Knowledge Space

Obtaining associated concepts for a given culture using Culturescope enables us to examine concepts encoded within the model parametric space. Beyond analyzing concepts individually, comparing the association strengths of concepts shared across cultures provides insight into how cultural knowledge is distributed within the model. To this end, we propose a cultural imbalance score that quantifies imbalances in relationships among cultures within the model's internal representation space.

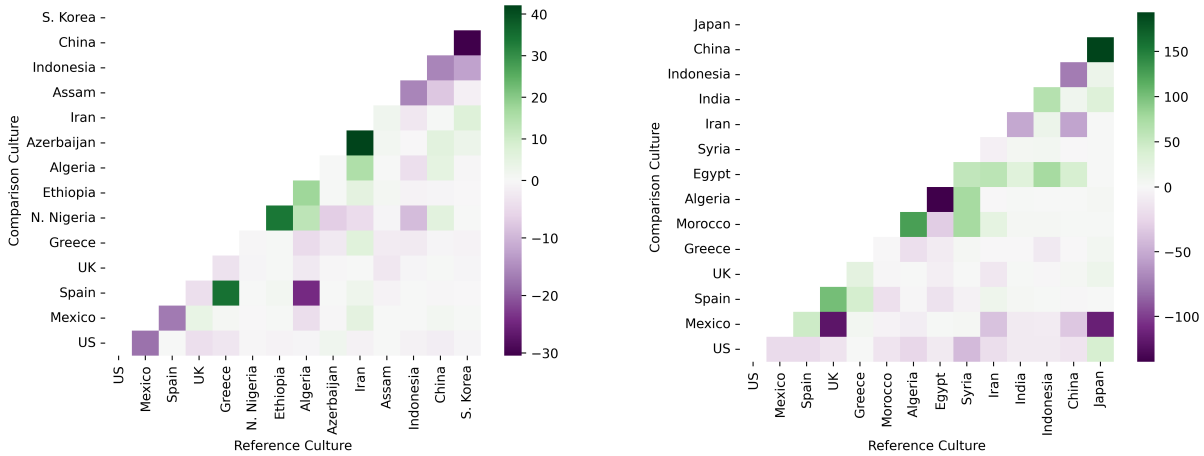
5.1 Cultural Imbalance score

For a pair of cultures consisting of a reference culture c_r and a comparison culture c_g, we first identify the set of overlapping concepts between the two cultures at layer l, denoted as $\mathcal{K}^{(l)}_{r,g}$. For each overlapping concept $k \in \mathcal{K}^{(l)}_{r,g}$, we compute a relative association score defined as the difference between the association score of the reference culture and that of the comparison culture:

$$\Delta^l_k(c_r, c_g) = s(k^l_j, c_r) - s(k^l_j, c_g), \qquad (4)$$

where $s(k^l_j, c)$ denotes the association score of concept k for culture c at layer l. A positive $\Delta^l_k(c_r, c_g)$ means that the overlapping concept k^l_j appears more often in c_r than in c_g.

The cultural imbalance score between the two cultures at layer l is then defined as the average relative association score across all overlapping concepts:



(a) BLEND, `Llama-3.1`

(b) CAMEL-2, `Llama-3.1`

Figure 4: We show heatmaps of CI scores between two cultures. If a pair of a reference culture and a comparison culture is a green-coloured cell, it indicates that the overlapping concepts are more strongly associated with the reference culture. On the other hand, if the cell is purple-coloured, it indicates that the overlapping concepts are more strongly associated with a comparison culture.

$$I^l(c_r, c_g) = \frac{1}{|\mathcal{K}^l_{r,g}|} \sum_{k \in \mathcal{K}^l_{r,g}} \Big(s(k^l_j, c_r) - s(k^l_j, c_g) \Big) . \tag{5}$$

A negative CI score for a culture indicates that overlapping concepts are less strongly associated with that culture than with others, suggesting a relatively weaker representation. In contrast, a positive CI score is more ambiguous: 1) it may reflect strong associations with culturally distinctive concepts, 2) strong and appropriate associations with shared concepts, or 3) over-association where shared concepts are disproportionately attributed to the target culture. As a result, positive CI scores do not necessarily imply culturally accurate representations.

5.2 CI Score Results

Figure 4 shows the CI scores for pairs of reference and comparison cultures averaged across layers. When identifying overlapping concepts, we consider the top-1000 concepts for BLEND and the top-500 concepts for CAMEL-2. These thresholds are determined empirically by examining how different values of λ affect the stability of the CI scores. Detailed results for alternative thresholds are provided in Appendix C.

For visualization clarity, we display only the lower triangle of the heatmap. A green-coloured cell indicates that the reference culture exhibits stronger association scores for overlapping concepts than the comparison culture. On the other hand, a purple-coloured cell indicates that the comparison culture exhibits stronger association scores than the reference culture. Ideally, a model should retain culture-specific concepts while accommodating shared ones, leading to positive CI scores (green-coloured cells).

From Figure 4, we observe darker cells between neighbouring or similar cultures, indicating a more pronounced imbalance in their concept associations. For example, in Figure 4a, Algeria shows positive relative association scores when compared with Northern Nigeria and Ethiopia. Similarly, in Figure 4b, the UK and Japan exhibit positive CI scores when compared with Spain and China, respectively. This indicates that the model more strongly associates overlapping concepts with the UK and Japan, either because shared concepts are more strongly attributed to these cultures or because some culturally distinctive concepts from Spain and China are misattributed to them. *More broadly, these patterns suggest that certain cultures exhibit weaker associations with shared concepts than others, analogous to cultural flattening in the real world.* We observe similar patterns across other models, which are reported in Appendix D.

6 Bridging Model Internals to Model Outputs

In §5.1, we identify imbalances in the cultural knowledge space represented within the model's internal representations. In this section, we further

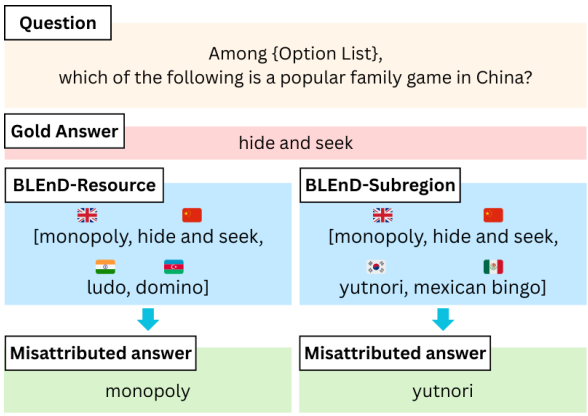


Figure 5: A cultural question about the popular family game in China from the BLEnD dataset (Myung et al., 2024). For the given question about China, if an LLM answers 'Yutnori', a popular family game in South Korea, it is caused by the effect of misattribution arising from confusing South Korea and China. On the other hand, if the answer is 'Monopoly', the LLM is generating an answer from a high-resource culture.

study how intrinsic imbalances influence model behaviour during downstream tasks. To investigate this connection, we construct a challenging evaluation setup that amplifies the overlap between cultures within model representations. Specifically, we first build a multiple-choice question (MCQ) dataset with hard negative options using BLEnD. We then correlate model performance on this dataset and the imbalance measured by Culturescope (§4.1). This setup allows us to examine whether internal cultural imbalances correlate with increased confusion when models must distinguish between the correct answer and a distracting answer.

6.1 Cultural MCQ with hard negatives

Figure 5 illustrates an example of option lists with hard negatives for a question about a popular family game in China. The inclusion of culturally plausible hard negative options targets misattributions that may arise from overlapping cultural knowledge and requires models to distinguish cultural nuances.

To amplify the model confusion that may arise across different cultural contexts, we categorize the 14 cultures in BLEnD along two dimensions: resource level and geographical subregion. For the resource dimension, we adopt the taxonomy proposed by Joshi et al. (2020), which we simplify into three levels: high-, mid-, and low-resource. For the region dimension, we group cultures into six subregional clusters. Additional details about these groupings are provided in Appendix E.1.

Using these groupings, we construct two evaluation partitions: BLEND-RESOURCE and BLEND-SUBREGION. Each partition uses a different strategy for generating hard negative options. We elaborate on the option creation in Appendix E.2.

	Acc.	Pearson's r	# of Cultures
		BLEND-RESOURCE	
Llama-3.1	0.4182	0.0206	7
aya-expanse	0.393	0.5366	5
Qwen2.5	0.4516	0.407	10
		BLEND-SUBREGION	
Llama-3.1	0.4225	0.218	7
aya-expanse	0.3531	0.5746	5
Qwen2.5	0.4479	0.4213	10

Table 2: We present accuracies of models on BLEND-RESOURCE and BLEND-SUBREGION. Correlation is calculated between each culture's accuracy and its CI score. We expect the CI score to become more negative as accuracy decreases, resulting in positive correlation values.

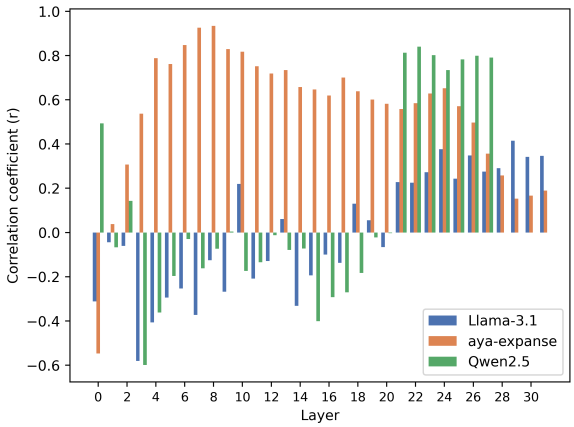


Figure 6: We present how the correlation value between CI scores and model accuracies changes across layers for BLEND-RESOURCE.

6.2 Correlation Results between CI scores and Model Performances

To see the relationship between CI scores, an intrinsic measure of cultural knowledge, and model performance, we compute Pearson's r between CI scores and model accuracies. Because CI scores are computed pairwise between cultures, we aggregate them to obtain a single CI score per culture. For each culture, we average its CI scores across all comparison cultures at each layer, and then average the resulting values across layers. The correlation

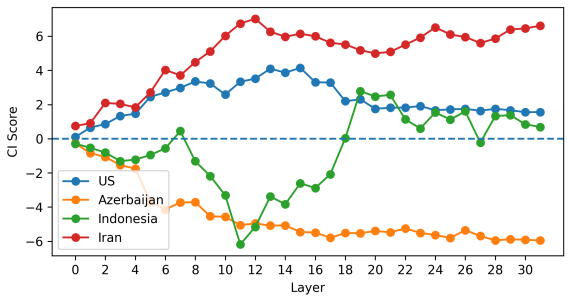


Figure 7: We present how the CI score changes across layers on the US, Azerbaijan, Indonesia, and Iran for BLEND on `Llama-3.1`.

analysis examines whether a lack of high association of shared concepts with respect to other cultures (resulting in a negative CI score) is associated with poorer performance on questions targeting that culture. As positive CI scores do not necessarily only imply stronger cultural associations of shared concepts (§5.1), we only consider cultures with negative CI scores for the computation (# of Cultures in Table 2).

Table 2 reports the correlation between the averaged CI scores and accuracies for cultures. Across all models, we observe a moderately positive relationship between the two, reflecting that minority cultures with low averaged CI scores perform worse on questions pertaining to that culture. The correlation scores are between 0.2-0.57, with the exception of `Llama-3.1` on BLEND-RESOURCE. When the correlations are calculated per layer, as shown in Figure 6, the scores are often much stronger than those obtained using the averaged CI scores, especially with deeper layers exhibiting positive correlations in the 0.25-0.85 range, including `Llama-3.1`. A layer-wise correlation analysis on BLEND-SUBREGION, which exhibits a similar trend, is shown in Appendix F.1. We report p-values from correlation tests in Appendix F.2.

Plotting the CI scores across layers, in Figure 7, we observe that differences in CI scores across cultures become more pronounced in deeper layers. The US and Iran exhibit rising positive CI scores, whereas Azerbaijan and Indonesia move toward negative CI scores. This implies that stronger correlations observed in deeper layers are associated with the increasing distinction in CI scores across cultures. *This suggests that cultural knowledge is processed differently across model layers, and that representations in specific layers may better explain model performance.*

7 Discussion & Conclusion

Our analysis shows that deeper layers produce fewer concepts but with higher association scores. This suggests that later layers may generate concepts that are more strongly associated with specific cultures, while earlier layers produce a larger number of candidate concepts with weaker cultural associations. One possible interpretation is that earlier layers enrich representations by activating a broad set of related concepts, whereas deeper layers selectively emphasize concepts that are most relevant to the context.

This interpretation is consistent with recent findings on the role of model layers, which suggest that earlier layers tend to enrich subject representations, while middle and later layers focus on selecting task-relevant attributes (Geva et al., 2023; Farahani and Johansson, 2024; Yu et al., 2024b). In this view, early layers may populate a broader cultural concept space, while deeper layers perform a selection process that surfaces a smaller set of culturally associated concepts. Our results suggest that cultural knowledge processing within language models may follow a similar pattern. However, we also observe that cultural imbalance scores tend to increase in later layers, which also correlate with model performances. This pattern indicates that the selection process in deeper layers amplifies asymmetries in cultural representations, potentially leading to phenomena such as cultural flattening or misattribution.

Our finding of this trend in later layers holds several implications for better aligning LLMs culturally. First, as cultural imbalance emerges more strongly in later layers, interventions aimed at improving cultural representation may benefit from focusing on these layers. For example, steering or editing methods could prioritize layers that exhibit the largest imbalance signals. Second, for cultures that consistently exhibit negative CI scores, targeted fine-tuning or data augmentation may help strengthen the model's associations with culturally specific concepts. Finally, the CI metric introduced in this work provides a tool for monitoring how cultural associations evolve during post-training or alignment procedures, enabling practitioners to evaluate whether interventions improve or worsen cultural balance within the model's internal representations.

Limitations

Dataset Related Although a growing number of benchmarks evaluate cultural knowledge in LLMs, few are suitable for our setup. To compare cultural knowledge across cultures, datasets must maintain a consistent QA format. As a result, our analyses inherit limitations of existing benchmarks, such as limited coverage of fine-grained cultures or subcultures. Nevertheless, our method is model- and task-agnostic and can be applied to any dataset that satisfies these requirements.

Our definition of low-resource cultures follows the approximation based on the most-spoken language from prior work (Joshi et al., 2020). In addition, the construction of MCQs with hard negatives (§6.1) involves random sampling. While this introduces some variability, it is a practical compromise given the cost of evaluating all possible combinations. We mitigate this by focusing on aggregate trends rather than individual instances.

Culturescope Related Due to computational constraints, our experiments are limited to models of approximately 8B parameters. In addition, because Culturescope is reference-free, the extracted concepts may contain noise.

Another limitation concerns the interpretation of positive CI scores. While they indicate stronger associations between concepts and a culture, they do not necessarily imply culturally accurate knowledge. Some associations may arise if concepts from other cultures are subsumed under a target culture during later processing. Distinguishing these cases would require reliable ground-truth cultural labels for concepts, which remains a challenging annotation task.

Finally, our analysis does not track how concepts activated in early layers propagate to later layers and treats each layer as a single unit without distinguishing attention and MLP sublayers. Future work could examine how early representations influence later concept selection and how different transformer components contribute to the processing of cultural knowledge.

Miscellaneous An AI writing assistant tool is used for a grammar check and stylistic improvement. None of the AI generations is used as is.

References

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.

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.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA. Association for Computational Linguistics.

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.

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.

Mehrdad Farahani and Richard Johansson. 2024. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16966–16977, Miami, Florida, USA. Association for Computational Linguistics.

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.

Robert J. Lieber and Ruth Weisberg. 2003. Globalization, culture, and identities in crisis. *International Journal of Politics, Culture, and Society*, 16(2):273–296.

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. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

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.

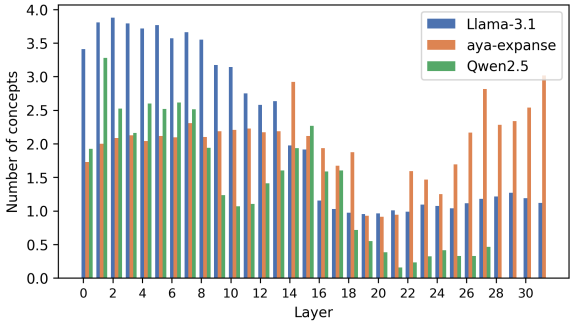
Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: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

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.

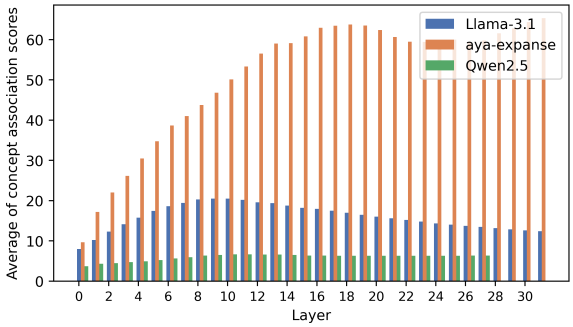
Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. 2024a. Revealing the parametric knowledge of language models: A unified framework for attribution methods. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8173–8186, Bangkok,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

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024b. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA. Association for Computational Linguistics.

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.



(a) Averaged number of concepts generated per layer



(b) Averaged association scores per layer

Figure 8: We show how Culturescope outputs differs across layers w.r.t. the number of generated concepts per question (Figure 8a) and the association scores for generated concepts (Figure 8b), for the CAMeL-2 dataset.

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 cultures to keep a similar culture distribution to the BLEnD dataset. We provide brief data statistics and characteristics of datasets in Table 4.

B Culturescope

B.1 Implementation Details

In Figure 9, we present the inspection prompt we use during the Patching stage of Culturescope.

	Culturescope	Irrelelevant
	BLEnD	
`Llama-3.1`	0.2360 ±0.1975	0.1764 ±0.1044
`aya-expanse`	0.2751 ±0.2009	0.1737 ±0.0989
`Qwen2.5`	0.2748 ±0.1925	0.1755 ±0.103
	CAMeL-2	
`Llama-3.1`	0.1989 ±0.1513	0.1474 ±0.0939
`aya-expanse`	0.1776 ±0.1811	0.1253 ±0.0966
`Qwen2.5`	0.1364 ±0.1369	0.1088 ±0.1027

Table 3: We present the irrelevant patching results with the cosine similarity scores between generated concepts for each input and its gold answer for BLEnD and CAMeL-2. We report the mean and standard deviation of the similarity scores.

B.2 Evaluation of Culturescope

Overall, we observe that culturescope generates culturally grounded concepts that align with domains of the prompt, such as food, traditions, and sports (Table 10, Table 11). Figure 8 shows a similar trend with the BLEnD dataset, as deeper layers produce fewer concepts, the association scores increase.

B.3 Number of Unique Concepts Generated per Model

Table 5 describes the number of unique generated concepts for each dataset and model. We observe that `aya-expanse` generates fewer unique concepts compared to the other two models. We attribute `aya-expanse`'s higher association scores in general to this.

B.4 Comparison with Irrelevant Patching

In Table 3, we see higher cosine similarity of generated concepts and gold answers from our method, Culturescope. The standard deviation is high due to noisy concepts. However, as we only consider top-λ concepts and downweigh concepts through our concept scoring phase, we expect the impact of noisy concepts as minimal.

C Varying Threshold for finding an optimal top-k

Figure 10 shows how the degree of edge connectivity changes by varying λ. We define the degree of edge connectivity D as:

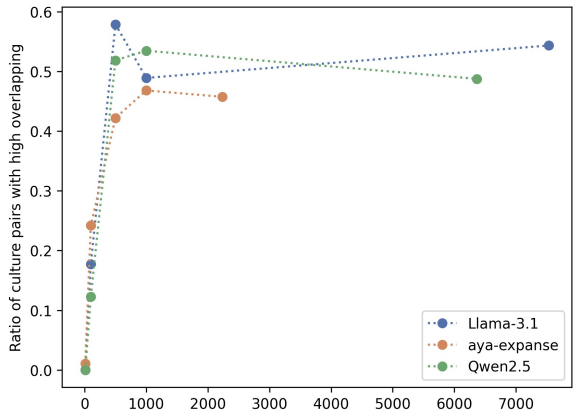
$$D = \frac{2 \cdot E}{|\mathcal{C}|(|\mathcal{C}| - 1)} \quad (6)$$

Dataset	Task	Number of Questions	Domain	List of Cultures
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

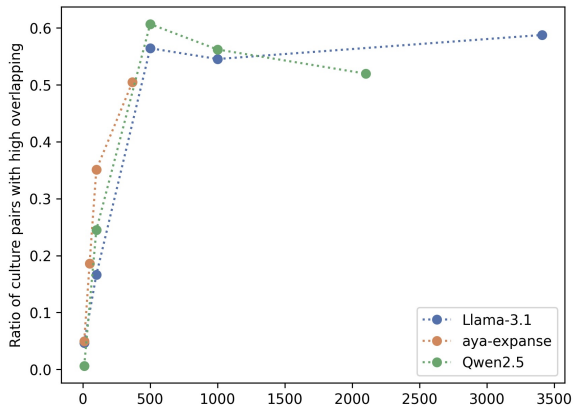
Table 4: Details for the datasets

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

Figure 9: An inspection prompt used during the Patching stage of Culturescope.



(a) Results for BLEnD.



(b) Results for CAMeL-2.

Figure 10: We show how the degree of edge connectivity changes by varying λ for both datasets. x-axis represents different λ.

where E is the number of culture pairs that have higher CI scores than the average CI scores across cultures. $|\mathcal{C}|$ denotes the total number of cultures in a dataset. We observe that D remains steady when λ is set to 1000 for BLEnD and 500 for CAMeL-2. Since `aya-expanse` generates fewer than 500 concepts per culture, we set λ to 100.

D CI Score Results

Figure 11 shows that similar trends hold for BLEnD with `aya-expanse` and `Qwen2.5`. For CAMeL-2, however, we observe overall positive CI scores from the reference cultures. Given that both models achieve higher accuracy on CAMeL-2, this pattern may indicate that they retrieve culturally specific knowledge more reliably.

E Details of Cultural MCQ with hard negatives

E.1 Cultural Groups

As mentioned in §6.1, our dataset focuses on 14 cultures 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 subregion dimension, we group cultures into six subregions. We split Asia into three subregions, which leaves us six subregional groups:

	Llama-3.1			aya-expanse			Qwen2.5		
	Min	Avg.	Max	Min	Avg.	Max	Min	Avg.	Max
BLEnD	4199	4502.57	5028	1317	1482.71	1739	3689	3891.71	4255
CAMeL-2	1852	2247.35	2751	227	392.21	641	1261	1470.85	1916

Table 5: We show the number of unique generated concepts for each dataset and model.

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 6 shows the culture entities that correspond to each group.

E.2 Creation of hard negatives

For BLEND-RESOURCE, given a question q targeting culture c, we sample one culture from each of the three resource levels excluding c. We then obtain the corresponding gold answers when each sampled culture is substituted into q in place of c. These answers form three hard negative options: o_{high}, o_{mid}, and o_{low}.

For BLEND-SUBREGION, we sample one culture from the same geographical region as c (excluding c) and extract its corresponding answer to construct a region-based hard negative option, denoted $o_{same\text{-}region}$. Two additional options ($o_{region1}$ and $o_{region2}$) are randomly selected from cultures in different regions. Finally, all options are shuffled to avoid positional bias.

E.3 Descriptive Statistics of Option Sampling

The creation of BLEND-RESOURCE and BLEND-SUBREGION involves random selection among possible options (§6.1). Table 7 shows the number of questions per group in both partitions. Table 8 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 a 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 cultures from three resource groups.

Dimension	Groups	Cultures
Resource	High	Algeria, China, Iran, Mexico, Spain, UK, US
	Mid	Greece, Indonesia, South Korea
	Low	Assam, Azerbaijan, Ethiopia, Northern Nigeria
Subregion	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 6: Culture groups and the list of cultures belong to each group

BLEND-RESOURCE		BLEND-SUBREGION	
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 7: Number of questions per group in Cultural MCQ with hard negatives

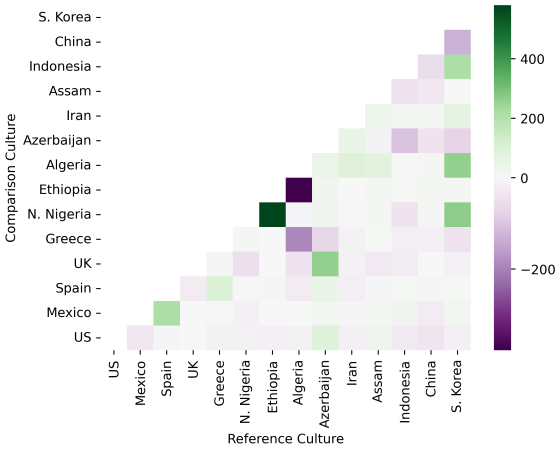
F Additional Correlation Tests

F.1 Layer-wise Correlation Results on BLEND-SUBREGION

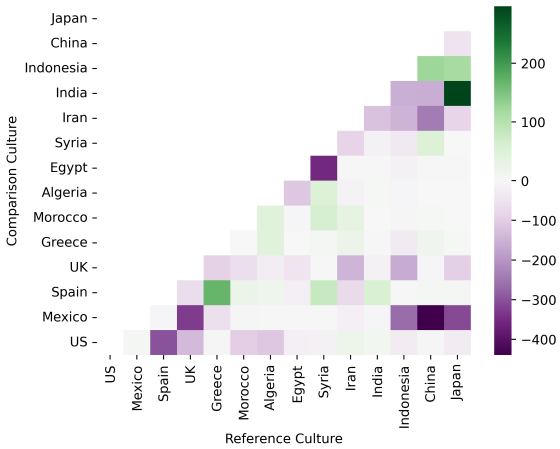
Figure 12 shows how the correlation value changes across layers for BLEND-SUBREGION. We observe that deeper layers correlate more strongly with model performance, as shown in Figure 6.

F.2 Statistical Significance of Correlation Tests

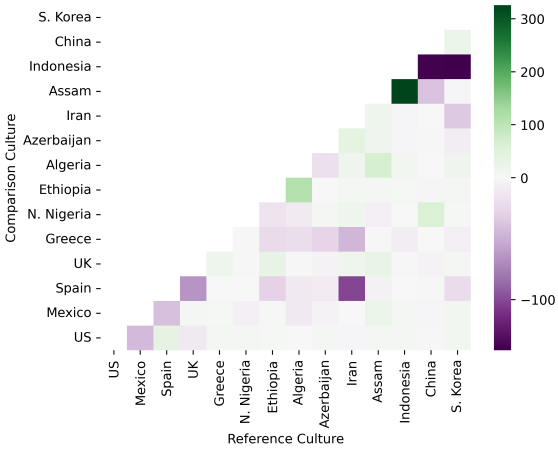
We present statistical significance of results in Table 2, Figure 6, and Figure 12 in Table 9. Columns with l show correlation results with layer-wise anal-



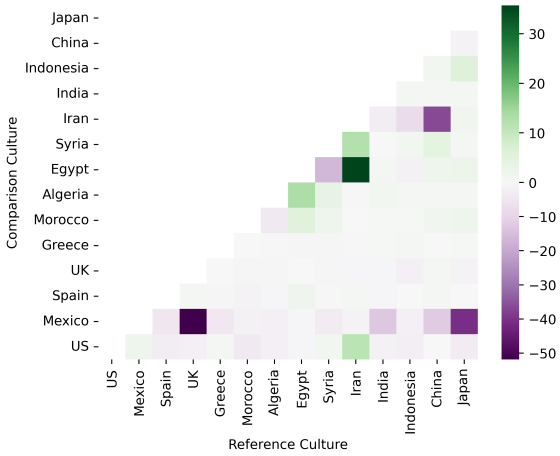
(a) BLEnD - aya-expanse



(b) CAMeL-2 - aya-expanse



(c) BLEnD - Qwen2.5



(d) CAMeL-2 - Qwen2.5

Figure 11: We present CI score results on aya-expanse Qwen2.5.

Option type	Mean ± s.d.	95% CI
BLEND-SUBREGION		
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 8: 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-SUBREGION. 'Others' means randomly selected options.

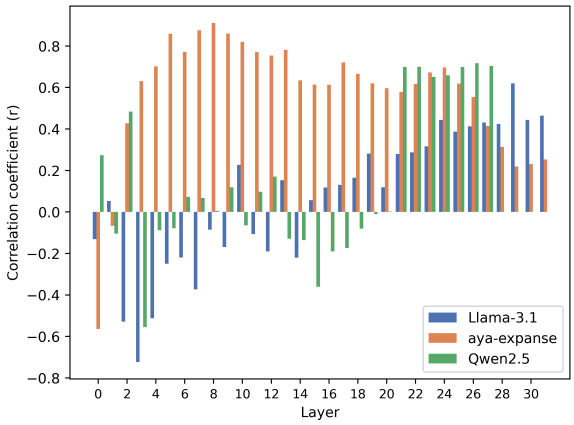


Figure 12: We present how the correlation value changes across layers for BLEND-SUBREGION.

	# of Cultures	Pearson's r	p-value	Pearson's r (l)	p-value (l)
		BLEND-RESOURCE			
Llama-3.1	7.18	0.0206	0.965	0.0658	0.3199
aya-expanse	5.37	0.5366	0.351	0.3516	0.000
Qwen2.5	8.9	0.407	0.2422	0.2266	0.000
		BLEND-SUBREGION			
Llama-3.1	7.18	0.218	0.6385	0.1249	0.0584
aya-expanse	5.37	0.5746	0.3109	0.3920	0.000
Qwen2.5	8.9	0.4213	0.2252	0.2587	0.000

Table 9: We present statistical significance of results in Table 2, Figure 6, and Figure 12.

yses. For layer-wise analyses, we treat each (accuracy, CI score) pair for a given culture as a single data point to ensure sufficient sample size. We observe high p-values in the averaged results due to the small sample size. However, p-values with layer-wise analyses are statistically significant, validating our correlation results.

Model	Culture	Top-10 concepts
Llama-3.1	Algeria	Moroccan, Algerian, Couscous, Algiers, Tunisia, Berber, Algeria, Tunis, Libya, Tunisian
	Greece	Souvlaki, Greece, Socrates, Plato, Acropolis, Aristotle, Gyro, Santorini, Athens, Crete
	Ethiopia	Injera, Ethiopia, Djibouti, Eritrea, Addis Ababa, Berbere, Ethiopian, Somalia, Tibs, Teff
	Spain	Tapas, Paella, Painter, More Rows, Gazpacho, Seville, Catalonia, Valencia, Flamenco, Sangria
	Azerbaijan	Baku, Quizzed, Algérie, Pilaf, Ukraine, Tbilisi, Aljazair, Belarus, Azerbaijan, Ended
	Assam	Assam, Krishna, Delhi, Ganges, Mumbai, Arunachal Pradesh, Wong, Rama, Pongal, Rajasthan
	Indonesia	Indonesia, Jakarta, Bali, Bandung, Surabaya, Brunei, Sate, Indonesian, Yogyakarta, Nasi Goreng
	Iran	Tehran, Fesenjan, Iranian, Persepolis, Persian, Abgusht, Nowruz, Tajikistan, Tahdig, Isfahan
	US	Nyc, Magic Kingdom, Boston, Green Bay Packers, Chicago, San Francisco, Miami, Yankees, Hut, Mickey Mouse
	China	Guangzhou, Tianjin, Shenzhen, Dumplings, Nanjing, Dragon, Szechuan, Dim Sum, Xinjiang, Shanghai
	UK	Wales, Scotland, Edinburgh, Northern Ireland, Ireland, Buckingham Palace, Great Britain, Flags, Glasgow, Royal Family
aya-expanse	Algeria	Tripoli, Morocco, Tunisia, Tunis, Algeria, Libya, Couscous, North Africa, Cairo, Egypt
	Greece	Greek, Athens, Santorini, Greece, Mykonos, Gyro, Zeus, Greek Mythology, Poseidon, Albania
	Ethiopia	Ethiopia, Somalia, Injera Bread, Injera, Uganda, Sudan, Tanzania, Waraa, Nile, African
	Spain	Tapas, Paella, Valencia, Jenga, Go, Seville, Bilbao, Barcelona, La Liga, Granada
	Azerbaijan	Georgia, Armenia, Azerbaijan, Moa, Belarus, Ukraine, Intaa, Kazakhstan, Kyrgyzstan, Azimab
	Assam	Nepal, Jateen, Bhutan, Sri Lanka, Bangladesh, Kolkata, Mumbai, Chennai, Delhi, India
	Indonesia	Indonesia, Gado, Malaysia, Vietnam, Thailand, Jakarta, Philippines, Bali, Nasi Goreng, Gadoid
	Iran	Tap, Ghormeh Sabzi, Tehran, Valley, Baghdad, Mehri, , Nowruz, Persian New Year, Ali Baba
	US	, New England Patriots, Tom Brady, Tom, Nfl, Colorado, Dell, Hp, Massachusetts, Boston
	China	, Macau, Shanghai, Beijing, Guangzhou, Shenzhen, Roger Federer, Hong Kong, Taiwan, Bruce Lee
	UK	Wales, Scotland, Great Britain, Cambridge, None, Northern Ireland, Chelsea, Liverpool, Big Ben, Manchester
Qwen2.5	Algeria	Casablanca, Algiers, Tunis, Tunisia, Moroccan, Gla, Algeria, Morocco, Tangier, North Africa
	Greece	Macedonia, Olympia, Parthenon, Souvlaki, Acropolis, Greece, Greek, Ouzo, Raki, Kostas
	Ethiopia	Era, Injera, Kitfo, East Africa, Doro Wat, Eritrea, Ethiopia, Ethiopian Cuisine, Provided, Djibouti
	Spain	Siege, Valencia, Basque Country, Paella, Gazpacho, Catalonia, Galicia, Juventus, Andalusia, Tapas
	Azerbaijan	Baku, Azerbaijan, Konyol, Georgia, Manty, Tbilisi, Caucasus, Armenia, Kazakhstan, Uzbekistan
	Assam	Sri Lanka, Bangladesh, Pongal, Assam, Meghalaya, Bhutan, Kolkata, Northeast India, Chennai, Mumbai
	Indonesia	Indonesia, Kuala Lumpur, Bandung, Brunei, Malaysia, Hanoi, Sumatra, Rendang, Satay, Borneo
	Iran	Isfahan, Fesenjan, Red Sea, Shiraz, Persian Cuisine, Good Friday, Iranian Cuisine, Persian New Year, Mashhad, Hindu Kush
	US	Texas, Lsat, Mcat, Los Angeles, Chicago, Gre, Dallas Cowboys, Boston, Houston, Alaska
	China	Guangzhou, Nanjing, Dumplings, Shanghai, Guangdong, Hong Kong, Midautumn Festival, Beijing, Moon, Shenzhen
	UK	Wales, Northern Ireland, Manchester, Ireland, Leeds, England, Gordon Ramsay, Birmingham, London Eye, Battleship

Table 10: We show top-10 generated concepts from Culturescope on BLEnD.

Model	Culture	Top-10 concepts
Llama-3.1	Japan	Udon, Ramen, Kyoto, Mochi, Osaka, Tempura, Sushi, Wasabi, Miso Soup, Teriyaki
	Mexico	Taco, Salsa, Vanilla, Burrito, Guacamole, Cookie, Ice Cream, Chocolate, Sweet, Cake
	Morocco	Amaranteck, Marrakech, Fez, Rabat, Chefchaouen, 50, Fes, Algeria, Casablanca, Atlas Mountains
	Spain	Valencia, Catalonia, Granada, San Sebastian, Bilbao, Seville, Andalusia, Madrid, Toledo, Galicia
	Syria	Aleppo, Damascus, Homs, Daraa, Hama, Palmyra, War, Spanish Language, Aisha, Idlib
	UK	Haggis, Cristiano Ronaldo, Real Madrid, Chelsea, Goal, Liverpool, Sport, Lionel Messi, Player, Messi
	US	New York, Washington, Texas, Florida, Illinois, California, Seattle, America, North America, Canada
	Algeria	Mali, Chad, Nigeria, Niger, Mauritania, Morocco, Tunisia, Angola, Libya, Cameroon
	China	Dumpling, Egg, Bok Choy, Dumplings, Lasagna, Chinese Food, National Parks, Lake Tahoe, Vegan, Fried
	Egypt	Nile River, Nile, Na, Alexandria, Sphinx, Pyramids, Suez Canal, Giza, Nubia, Aswan
	Greece	Greece, Macedonia, Athens, Albania, Crete, Bulgaria, Aegean Sea, Peloponnese, Aegean, Thessaly
	India	Idli, Vada, Dosa, Cannabis, Marijuana, Curry, Chicken Tikka Masala, Dal, Naan, Carnaval
	Indonesia	False, Jackfruit, Word Embeddings, Shahnameh, Machine Learning, Durian, Nlp, Persian Art, Prophet Muhammad, File
	Iran	Ayatollah, Mullah, Islamic, Worship, Minaret, Islam, Kaaba, Dome, Holy Quran, Imran Khan
aya-expanse	Japan	Algae, Yeast, Fungus, Sushi, Ramen, Japanese, Tempura, North Korea, Mold, Fungal
	Mexico	Candy, Snack, Oaxaca, Sweet, Dessert, Peru, Starbucks, Risotto, Uruguay, Sweets
	Morocco	Control, Algeria, Tunisia, Libya, Arabs, North Africa, Morocco, Tourism, Egypt, Hotel
	Spain	Painting, Valencia, Photography, Barcelona, Sculpture, Artist, Synagogue, Temple, Art Gallery, Seville
	Syria	Adh, Iraq, Alfaisal, Amu, Iran, Ken, West Asia, Star, Cnn, Bbc
	UK	Potatos, Scotland, Wales, Northern Ireland, Sandwich, Manchester United, Premier League, Gingerbread, Bubble Bath, Champagne
	US	North America, Canada, United States, Mexico, Europe, Asia, Florida, Houston, Dallas, Washington
	Algeria	Chad, Congo, Elan, Elan Pharmaceuticals, Elan Corporation Plc, Cameroon, Cape Verde, Central African Republic, Comoros, Côte Divoire
	China	Chura, Shanghai, Shenzhen, Guangzhou, Beijing, Mallard Anas Platyrhynchos, Valentine, China, Platinum, Singapore
	Egypt	Alsaud, Alsisi, Text, Aljazeera Tv Satellite, New York City, Aljazeera Network, Aljazeera Satellite, Aljazeera Tv, Aljazeera Tv Network, Article
	Greece	Macedonia, Bulgaria, Albania, Athens, Greece, Cyprus, Mediterranean, Greek, Serbia, History
	India	Legalization, Pot, Cannabis, Drug, Marijuana, Legal, Weight Loss, Garcinia, Curry, Fat Burner
	Indonesia	Gado, Spices, Gado Gado, Falafel, Hummus, Baklawa, Na, Herbs, Volleyball, Baklava
	Iran	Shiraz, Tehran, List, Classification, Prayer, Sunni, Prophet Muhammad, Esfahan, Isfahan, Quran
Qwen2.5	Japan	Mochi Ice Cream, Osaka, Lobster, Kyoto, Coordinates, Shrimp, Edo, Seoul, Koyaa, Korea
	Mexico	Candy, Sweet, Ice Cream, Cake, Dessert, Cookies, Cookie, Alexander The Great, Euro, Strawberry
	Morocco	Gaziantep, Ail, Ase, Tunisia, Algeria, Mauritania, Municipality, Lufthansa, Sudan, Air France
	Spain	Alhambra, Valencia, Madrid, Granada, Location West Asia, Alcazaba, Barcelona, Palace, Western World, Caribbean
	Syria	Legislature, Product Category Electronics, Boko Haram, Parliament, Bbc, Political, Cnn, Fox News, Democracy, Taliban
	UK	Scotland, Vodka, Whisky, Beer, Alcohol, England, Lancashire, Wine, Wales, Dairy
	US	Tree, Los Angeles, Clinton, Joy, Trump, Happiness, Amusement, Obama, Pond, Election
	Algeria	Katrina Kaif, And Electronics, Nonsense, Famous Actors, And Electronics Companies, Rail, Central America, Names, Taxi, Amal
	China	Guangzhou, Fan Bingbing, Nanjing, Shanghai, Beijing, Zhang Ziyi, Yao Ming, Guangdong, Macau, Taiwan
	Egypt	Boston, New Jersey, Philadelphia, Fox News, Windows, Taliban, Linux, Cnn, Bbc, Al Arabiya
	Greece	Tennis, Parthenon, Ancient Greece, Acropolis, Indus Valley, Sumer, Greece, Babylon, Basketball, Ancient Rome
	India	Marijuana, Cannabis, Weight Loss, Poppy, Drug, Opium, Diet Pill, Qatari, Morphine, Pain Relief
	Indonesia	Great Noor Mosque, Listed, Paradise, British Empire, Topics, Description, Mongolia, Ruins, Roman Empire, Dumpling
	Iran	Sufi, Yazidi, Alawite, Sunni, Druze, Christian, Belief, Sect, Prayer, Temple

Table 11: We show top-10 generated concepts from Culturescope on CAMeL-2.