Deciphering Cultural Representations in Large Language Models via Sparse Autoencoders

Chenye Zou[1], Difan Jiao[2], Lijie Hu[1]
[1]Mohamed bin Zayed University of Artificial Intelligence, [2] University of Toronto
zoucy2001@gmail.com; lijie.hu@mbzuai.ac.ae

Abstract

Large Language Models (LLMs) are increasingly deployed worldwide, yet they exhibit strong Western-centric biases, and the internal mechanisms governing their cultural behaviors remain poorly understood. Prior work has identified so-called cultural neurons, but individual neurons are often polysemous, conflating abstract cultural knowledge with surface-level lexical cues due to superposition. We apply Sparse Autoencoders (SAEs) to decompose intermediate LLM activations into sparse, interpretable feature representations that disentangle these factors. This analysis reveals culturally selective features that remain invariant across paraphrasing and task formats, indicating abstraction beyond lexical correlations. Through targeted feature ablation, we provide causal evidence that these features are necessary for cultural reasoning: their removal selectively degrades performance on culturally conditioned tasks. Furthermore, we show that steering model activations along these feature directions is sufficient to systematically modulate cultural-related knowledge generation, without retraining. Together, our results offer the first causal evidence that LLMs encode cultural knowledge as decoupled semantic structures rather than surface patterns, enabling a scalable pathway toward cultural alignment through mechanistic intervention. Code is available at https://github.com/IAN-YE/Cultural-features-SAE.

1 Introduction

As large language models (LLMs) become integral to global communication and decision-making (Cheng et al., 2025a,b; Yang et al., 2025), their ability to navigate culturally diverse contexts is increasingly critical (Cheng et al., 2024; Li et al., 2025b). However, the cultural knowledge encoded in current LLMs largely reflects biases inherited from predominantly Western-centric training data, leading to uneven cultural representation across regions and communities (Dodge et al., 2021; Wang et al., 2024). A growing body of work has shown that such imbalance manifests as systematic cultural biases in model behavior, evidenced by a wide range of cultural benchmarks and evaluations (Myung et al., 2024; Chiu et al., 2025; Rao et al., 2025). In response, numerous approaches have been proposed to improve cultural alignment or awareness in LLMs, often through data augmentation, prompting strategies, or post hoc adjustment methods (AlKhamissi et al., 2024; Seo et al., 2025; Zhang et al., 2025a).

Existing work has primarily approached cultural understanding in LLMs through causal interventions on representation spaces (Yu et al., 2025a) and neuron-level localization (Yamamoto et al., 2025; Su et al., 2025). Neuron-based methods aim to identify culture-specific neurons and analyze their influence on culturally conditioned model behavior. However, activation patching(Meng et al., 2022) and related causal tracing methods are often sensitive to prompt design and intervention location, making it difficult to isolate abstract cultural representations from task- or context-specific effects (Zhang and Nanda, 2023; Makelov et al., 2024). Moreover, neuron-based methods are sometimes unreliable, due to "superposition" (Elhage et al., 2021), which suggests that neural networks often consolidate multiple unrelated concepts into a single neuron. As such, it is important to use a more reliable and interpretable method to analyze cultures in LLMs. While these works use mutual information to study cultural representations, they stop short of demonstrating how these insights can be translated into effective interventions that improve performance on culture-related questions.

In this work, we adopt Sparse Autoencoders (SAEs) (Huben et al., 2024; Bricken et al., 2023; Cunningham et al., 2023) to decompose language model activations into sparse, interpretable feature

Findings of the Association for Computational Linguistics: ACL 2026, pages 5656–5677

directions, as they are uniquely suited for deciphering the complex nature of cultural representations. First, cultural knowledge in LLMs is often "entangled" within polysemantic neurons (Yamamoto et al., 2025) and SAEs allow us to disentangle these multi-faceted cultural identities into monosemantic features, effectively separating deep-seated cultural norms from superficial lexical associations. Second, the independent layer-wise training of SAEs provides a high-resolution map of the model's internal "cultural processing pipeline," enabling us to pinpoint exactly where abstract cultural reasoning emerges. Concretely, our layer-wise analysis reveals a sharp crossover transition in which cross-task similarity for the same culture begins to exceed cross-cultural similarity, showing that early layers prioritize task-format encoding while upper layers invert this hierarchy—a structural property invisible to neuron-level methods.

Given the advantages of SAEs, we present a mechanistic investigation of cultural representation using SAEs. By decomposing intermediate activations into a sparse set of interpretable features, we overcome the noise of token-level activations and the polysemanticity of raw neurons and finally increase the accuracy of cultural-related benchmarks. Our work progresses through three critical research questions: We investigate cultural representations in LLMs through three complementary questions. First, we examine whether LLMs form abstract cultural representations that are invariant to surface-level lexical forms (RQ1). We identify culture-selective features that activate consistently across diverse paraphrases, introduce a metric to quantify their cultural selectivity based on activation differences across cultures, and evaluate their stability across models, assessing whether cultural knowledge is organized in a disentangled semantic space rather than driven by lexical cues. Second, we test whether these features are causally necessary for cultural reasoning (RQ2) using directional ablation (Arditi et al., 2024), selectively suppressing culture-specific features during inference and observing corresponding performance degradation on culture-conditioned tasks. Strikingly, removing only the top-10 features within a critical layer collapses accuracy from 0.57 to 0.31, while matched random ablation causes negligible degradation—pinpointing a localized causal bottleneck where cultural information is concentrated. Finally, we examine causal sufficiency (RQ3) by applying activation steering to these features, demonstrating that targeted manipulation of internal representations can enhance culturally grounded model behavior without retraining.

In summary, our work makes three key contributions:

- **Mechanistic Identification:** We leverage Sparse Autoencoders (SAEs) to disentangle and identify cultural features that capture abstract semantics beyond superficial lexical patterns and propose a metric to measure the cultures of SAE features.
- **Causal Validation:** We provide causal evidence that these representations are functionally necessary for the model's cultural reasoning, through targeted feature ablation.
- **Interpretable Control:** We demonstrate that activation steering along these features allows for precise, retraining-free modulation of cultural behavior and bias mitigation.

2 Related Work

Cultural Knowledge and Bias in LLMs Large language models (LLMs) exhibit social and cultural biases measurable by benchmarks such as StereoSet (Nadeem et al., 2021), CrowS-Pairs (Nangia et al., 2020), and HolisticBias (Smith et al., 2022). Recent efforts extend this focus to cultural knowledge, with CulturalBench (Chiu et al., 2025), BLEnD (Myung et al., 2024), and CultureBank (Shi et al., 2024) evaluating or curating region-specific cultural content. Broader multilingual and geographic assessments such as Global MMLU (Singh et al., 2025) and GeoMLAMA (Yin et al., 2022) reveal systematic performance gaps across cultures and languages. Studies of value alignment show Western-centric defaults in model behavior (Tao et al., 2024; Masoud et al., 2025), while cultural prompting improves cross-cultural consistency. Mitigation approaches range from culturally grounded data augmentation to self-debiased inference (Schick et al., 2021; Gallegos et al., 2025) and joint bias–hallucination control (Lin et al., 2024). Overall, existing work underscores persistent cultural asymmetries in LLMs and motivates culturally aware evaluation and alignment frameworks.

Mechanistic Interpretability Mechanistic interpretability aims to uncover how internal components of large language models (LLMs) implement

behavior (Wu et al., 2024). Foundational analyses identified attention circuits and induction heads that encode algorithmic patterns within transformers (Olah et al., 2020; Elhage et al., 2021; Zhang et al., 2025b; Zhang et al.; Jiao et al., 2026; Wang et al., 2025). Techniques such as activation patching and causal tracing localize causal features and pathways (Wang et al., 2022), while model-editing approaches like ROME and MEMIT, etc., modify factual associations by directly intervening in representations (Meng et al., 2022, 2023; Zhang et al., 2025c; Li et al., 2025a; Jiang et al., 2025; Yu et al., 2025b). Recent work leverages Sparse Autoencoders (SAEs) to uncover interpretable, disentangled features in LLM activations (Bricken et al., 2023; Templeton et al., 2024; Sun et al., 2026), enabling scalable decomposition of high-dimensional representations into semantically meaningful directions. SAEs support both human-interpretable visualization and automatic circuit discovery when combined with activation patching or causal probing (Geiger et al., 2025; Cunningham et al., 2023; Yao et al., 2025). Despite this progress, the relationship between learned sparse features and emergent model capabilities remains imperfectly understood, motivating ongoing work on hybrid causal–representation frameworks that link internal structure to high-level semantics.

3 Preliminary

In this work, we study how large language models (LLMs) internally represent and retrieve cultural knowledge. Before presenting our main experiments, we briefly introduce Sparse Autoencoders (SAEs) and steering LLMs as the core methodological tool for our causal and representational analyses, followed by a description of the datasets and model configurations used in our experiments.

3.1 Sparse Autoencoders (SAEs)

Sparse Autoencoders (SAEs) (Huben et al., 2024; Bricken et al., 2023) aim to decompose the dense, polysemantic activations of an LM into a sparse set of interpretable features. Given an intermediate activation $a(x) \in \mathbb{R}^d$ from a model's residual stream, a typical SAE consists of an encoder and a decoder. The encoder maps $a(x)$ into a higher-dimensional latent space $\mathbb{R}^k$ (where $k \gg d$):

$$f_s(x) = \text{ReLU}(W_{\text{enc}} a(x) + b_{\text{enc}}) \quad (1)$$

where $W_{\text{enc}} \in \mathbb{R}^{k \times d}$ and $b_{\text{enc}} \in \mathbb{R}^k$. The decoder then attempts to reconstruct the original activation from this sparse representation:

$$\hat{a}(x) = W_{\text{dec}} f_s(x) + b_{\text{dec}} \quad (2)$$

The model is trained by minimizing a loss function $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{sparsity}}$, where $\mathcal{L}_{\text{rec}} = \|a(x) - \hat{a}(x)\|_2^2$ ensures reconstruction fidelity, and the L_1 penalty on $f_s(x)$ encourages the activation vector to be sparse. Each dimension (feature) in $f_s(x)$ ideally corresponds to a single, monosemantic concept, making it a suitable lever for model steering.

3.2 Steering Large Language Models

Steering is defined as influencing an LLM's output toward a target concept while maintaining generation quality and coherence. Unlike traditional fine-tuning, steering seeks a minimal change to the model's computation by intervening in its internal representations. Formally, given a model M and a prompt x, we obtain a steered text $\tilde{y}$ by applying an intervention $\Phi(\cdot)$ on an intermediate representation h:

$$\tilde{y} = M_{h \leftarrow \Phi(h)}(x) \quad (3)$$

Recent work utilizes SAEs to achieve interpretable control (Deng et al., 2025). To steer an LLM toward a concept encoded in an SAE feature f_i at layer l, we first pass the latent representation x^l through the SAE encoder to obtain the activations a. Following Templeton et al. (2024), we define the intervention Φ by modifying the activation vector as:

$$\tilde{a}_j = \begin{cases} a_j + s \cdot a_{\max} & j = i \\ a_j & j \neq i \end{cases} \quad (4)$$

where s is the steering factor and $a_{\max}$ is the maximum activation of the feature. The steered representation is reconstructed via the SAE decoder as $\Phi(x^l) = W_{\text{dec}} \tilde{a} + b_{\text{dec}}$ before continuing the forward pass. Success is typically measured by the frequency of concept-related tokens in the generated text and the resulting perplexity.

3.3 Models

We incorporate a diverse set of LLMs along with their corresponding SAEs to ensure the robustness of our findings. We did not train custom SAEs for this work, we use SAEs from Gemma Scope Gemma-2 9B (Lieberum et al., 2024)[1], Llama Scope for Llama-3.1-8B (Lieberum et al., 2024).

[1]Here we exclude Gemma-2-2B, as its performance on MCQ tasks is marginal, with accuracy only around 25%.

3.4 CPC Dataset

We construct our dataset from four complementary sources: WVS (World Values Survey, 2022), GAS (Pew Global Attitudes Survey, 2022), NormAd (Rao et al., 2025) and BLEND (Myung et al., 2025). WVS provides cross-national values and beliefs, GAS covers broad sociocultural attitudes, and BLEND offers everyday cultural knowledge. It consists 1,500 questions for each country.

To enable our analysis, we also convert all question-answering entries into declarative sentence form using GPT-4.1. Specifically, we transform items formatted as "Question P?" into "Statement S.", for example, converting "What is the capital of X?" into "The capital of X is Y." This critical transformation step constructs our Cultural Probe Corpus (CPC), which allows us to directly utilize these declarative sentences to calculate the Perplexity of target cultural knowledge and extract Sparse Autoencoder (SAE) feature activations for subsequent Cultural Specificity Metric (CSM) analysis. We also include MCQ-type questions following the setting (Yu et al., 2025a) in the final evaluation.

To test whether models generalize beyond surface forms, we also rewrite each question into five paraphrased variants using ChatGPT (prompt in Appendix A). A feature that activates consistently across semantically equivalent but lexically diverse surface forms provides stronger evidence of genuine cultural encoding, directly addressing RQ1.

Cultures (14)	
US(*US*), UK(*UK*), China(*CN*), Spain(*ES*), Mexico(*MX*), Indonesia(*ID*),South Korea(*KR*), Greece(*GR*), Iran(*IR*), Algeria(*DZ*), Azerbaijan(*AZ*), Assam(*AS*), Northern-Nigeria(*NG*), Ethiopia(*ET*)	
Prompt Type	**Prompt example**
Original	What is the most famous traditional sport in <country>?
Cloze	The most famous traditional sport in <country> is: __
Paraphrased(2/5)	What is the most renowned traditional sport in <country>? ; Of all traditional sports, the most famous in <country> is: __
MCQ Type	Question:What is the most famous traditional sport in the UK? Options:A. ping pong B. silat C. soccer D. wrestling

Table 1: Overview of the cultures and examples in CPC dataset

4 Cultural SAE Feature Characterization

In this section, we formalize the conceptual and computational setup used throughout the paper to understand how cultural information is internally represented in large language models.

4.1 Finding Cultural Specific Features

We first compute the mean activation of each latent feature across cultures. Let $D = \{D_{C_1}, \ldots, D_{C_K}\}$ denote datasets corresponding to $K = 14$ cultures, and let $f_s(x)$ denote the activation of the latent feature s (extracted via an SAE) for input x.

Culture-specific activation. For a target culture C, we compute the mean activation of feature s over prompts that explicitly include cultural information:

$$\mu_s^C = \frac{1}{|D_C|} \sum_{x \in D_C} f_s(x). \tag{5}$$

To control for question semantics, we additionally compute the mean activation on the same set of questions but without any explicit cultural information, isolating whether a feature's activation is genuinely triggered by cultural context rather than by topic-level or task-format cues:

$$\mu_s^{\text{w/o } C} = \frac{1}{|D_{\text{w/o } C}|} \sum_{x \in D_{\text{w/o } C}} f_s(x). \tag{6}$$

Cross-cultural baseline. To measure how selectively a feature activates for culture C relative to other cultures, we compute the average activation of feature s across all cultures except C:

$$\gamma_s^C = \frac{1}{|D \setminus \{D_C\}|} \sum_{D_I \in D \setminus \{D_C\}} \frac{1}{|D_I|} \sum_{x \in D_I} f_s(x). \tag{7}$$

Culture-specificity score. Inspired by (Deng et al., 2025), we define the **culture-specificity score** of feature s for culture C as:

$$\nu_s^C = \mu_s^C - \gamma_s^C, \tag{8}$$

which measures the extent to which feature s is selectively associated with culture C compared to other cultures.

Figure 1 reveals culture-specificity score across sparse features, where individual features exhibit strong preferential activation for particular cultural contexts which prove the correctness of culture-specificity score.

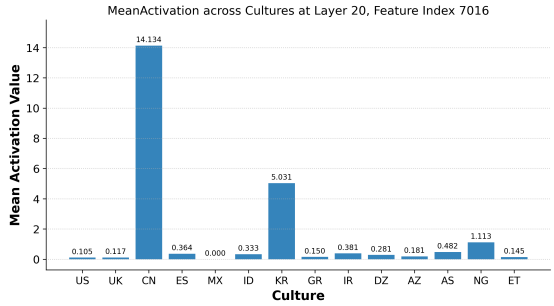


Figure 1: The mean activation of feature 7016 across different cultures in layer 20. The notably high mean activation in Chinese suggests that feature 7016 may be strongly associated with Chinese-specific knowledge.

Culture-injection effect. To quantify the extent to which a feature is directly induced by the presence of cultural information in the prompt, we define the **culture-injection effect** as:

$$\Delta_s^C = \mu_s^C - \mu_s^{\text{w/o } C}. \tag{9}$$

A large positive Δ_s^C indicates that feature s is strongly activated by explicit cultural cues, controlling for question semantics.

Joint identification of culture-selective features. Finally, we identify features that are both selectively associated with a target culture and directly induced by cultural information by jointly considering ν_s^C and Δ_s^C. In particular, features with large positive values of both scores are interpreted as *culture-selective representations* within the model.

4.2 Cross-Cultural Activation Analysis

We random to compute the culture-specificity score v for each culture. As shown in Figure 2a, the mean activation of the top-ranked SAE features is substantially higher than that of randomly selected features, whose activations remain close to zero. Across most cultures, we observe a sharp decay in mean activation among the highest-ranked features, with the top-ranked feature exhibiting a markedly larger activation than the remaining ones. In several cultures, the second-ranked feature also shows a noticeably elevated activation compared to lower-ranked features.

These observations indicate that a small number of highly ranked features dominate the culture-specific activation patterns, suggesting that such features capture concentrated and distinctive cultural signals.

4.3 Culture Injection Effects on Feature Activations

Figure 2a shows the culture injection effects at Layer 30, and other layers show similar results as shown in Appendix D. We observe that the magnitudes of the top activated features are notably similar across different cultures. In particular, the distributions of the top-k injection effects exhibit substantial overlap, with no culture consistently producing markedly larger activation shifts than others.

Despite this similarity in activation magnitudes, the specific feature indices that rank among the top activated features differ across cultures. This suggests that, at higher layers, cultural information induces activations of comparable strength, while remaining differentiated in terms of which latent features are engaged. It indicates that cultural prompts lead to a more uniform distribution of activation strengths across cultures, with cultural distinctions primarily reflected in feature identities rather than activation magnitudes.

5 Do LLMs Encode Abstract Cultural Concepts?

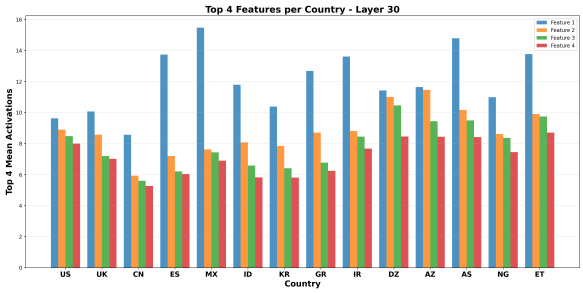
We investigate whether cultural information is explicitly encoded in the internal representations of LLMs, beyond surface-level lexical. Specifically, we ask whether culture remains recoverable from the model's internal states under substantial perturbations to prompt form and task structure.

To focus on the most selective and informative dimensions of the representation, we restrict our analysis to the top-4 SAE features ranked by our Cultural Salience Metric (CSM) for each culture. This stringent setting allows us to test whether cultural information is encoded in a compact and highly discriminative subset of the model's internal features.

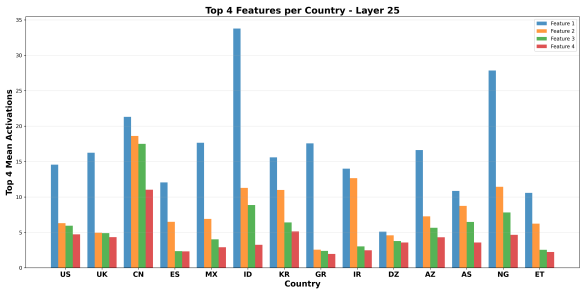
5.1 Robustness of Cultural Representations across Paraphrase Invariance

A necessary prerequisite for identifying abstract cultural representations is robustness to surface-level lexical variation. To rule out the possibility that culture-selective SAE features merely capture specific keywords or prompt templates, we first evaluate their invariance across paraphrased inputs that preserve the same underlying cultural semantics.

For each cultural statement in the CPC dataset,



(a) Distribution of top feature activation magnitudes induced by culture injection across cultures at Layer 30.



(b) Top activated features for a representative culture (China) at Layer 30 under culture injection.

Figure 2: Culture injection effects at Layer 30. (**Left**) Across different cultures, the magnitudes of the top activated SAE features exhibit highly similar distributions, indicating that culture injection induces activation shifts of comparable strength. (**Right**) For an individual culture, the most responsive features are sparse and localized to specific feature indices. Together, these results suggest that while the strength of culture-induced activations becomes more normalized at higher layers, cultural distinctions are primarily reflected in the selection of latent feature indices rather than activation magnitude.

we construct five paraphrased variants and extract SAE latent activations for each variant independently. We then identify the top-k culture-selective features (ranked by the Culture Specificity Metric) for each paraphrase and compute the average pairwise Jaccard similarity of these feature sets across paraphrases corresponding to the same statement.

We find that the overlap of top-ranked culture-selective features across paraphrases is consistently and significantly higher than that obtained from semantically unrelated statements drawn from different cultures. This result indicates that the identified features are not driven by specific lexical realizations or prompt templates, but instead respond to semantic content that remains stable under linguistic rephrasing.

Having established robustness to paraphrase-level perturbations, we next investigate whether these culture-selective features remain stable under more substantial changes in task structure and inference format.

5.2 Robustness of Cultural Representations across Task Formats

To assess whether these core cultural features encode task-invariant information or merely reflect prompt-specific artifacts, we compare SAE activation patterns between the Question (QA) and Cloze formats within our CPC corpus. Then we compute the Jaccard similarity of the top-5 features for each culture across formats, layer by layer, and compare it against a cross-cultural baseline.

As illustrated in Figure 3, we observe a pronounced representational crossover as information

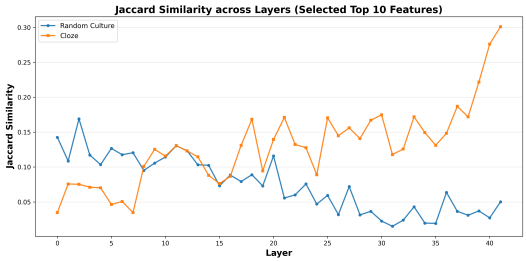


Figure 3: Layer-wise Jaccard similarity of the top-10 most culture-selective SAE features between Question (QA) and Cloze formats. The orange line shows cross-task similarity for the same culture, while the blue line denotes a cross-cultural baseline.

propagates through the transformer. In early and intermediate layers, the cross-task Jaccard similarity for the same culture remains consistently below the cross-cultural baseline, indicating that the most selective features at these layers are dominated by task-specific syntactic and token-level properties. Consequently, prompts from different cultures but sharing the same format appear more similar than prompts expressing the same culture across formats.

Around Layer 19, this trend reverses sharply: cross-task similarity begins to exceed the baseline and continues to increase in later layers. In the final layers, the same culture expressed in QA and Cloze formats shares approximately 25% of its top-5 features, while the baseline remains below 10%. Given the extreme sparsity of this analysis—considering only five features out of a 32,768-dimensional SAE space—a Jaccard similarity of this magnitude is highly non-trivial. These findings

demonstrate that LLMs encode cultural information as latent, task-invariant semantic features.

6 Causal Validation of Culture-Specific Features

In the previous section, we have identified culture-specific features that are closely related to abstract cultural concepts and demonstrate robust invariance across varied lexical forms and task structures. To evaluate the functional importance of the identified SAE latents, we perform a layer-wise causal ablation study across the residual stream of the model.

We specifically target the latents with the highest activation magnitudes, as these are hypothesized to be the primary encoders of cultural semantic information.

6.1 Model Interventions

Directional Ablation. Following prior work (Arditi et al., 2024; Ferrando et al., 2025), we analyze the causal contribution of a feature direction in the residual stream using *directional ablation*. Given a residual activation $\mathbf{x} \in \mathbb{R}^N$ and a feature direction $\mathbf{d} \in \mathbb{R}^N$, we remove the component of $\mathbf{x}$ aligned with $\mathbf{d}$ by subtracting its projection:

$$\mathbf{x}' = \mathbf{x} - \hat{\mathbf{d}}\hat{\mathbf{d}}^\top \mathbf{x}, \tag{10}$$

where $\hat{\mathbf{d}} = \frac{\mathbf{d}}{\|\mathbf{d}\|}$ denotes the unit vector. The ablated activation $\mathbf{x}'$ is then used in place of $\mathbf{x}$ for the remainder of the forward pass.

6.2 Ablation of Culture-Specific Features

Task Setup. We evaluate cultural reasoning using a Multiple-Choice Question (MCQ) format, where the model selects the culturally appropriate answer from a fixed candidate set. Model accuracy is measured as the proportion of correct cultural choices across our diagnostic dataset, and serves as the primary metric for quantifying the causal impact of feature ablation.

Following the methodology of Bricken et al. (2023), we zero-out the top-k (where $k \in 1, 5, 10$) features during the forward pass and measure the degradation in next-token prediction accuracy on our diagnostic dataset. We do not consider larger values of k because beyond the top 10 features, the mean activation values rapidly decrease and are typically below 1, making their contribution negligible for the purposes of this experiment.

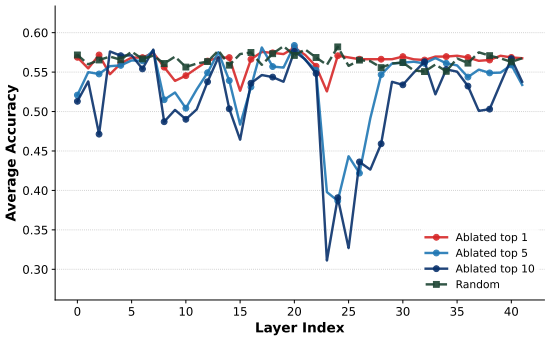


Figure 4: Layer-wise accuracy under top-k SAE feature ablation of Gemma2-9B. Random feature ablation with matched cardinality is shown as a negative control.

As illustrated in Figure 4, our results reveal a distinct hierarchical sensitivity to feature intervention: We observe a prominent sensitivity spike between Layer 22 and Layer 26. Specifically, at Layer 23, ablating the top-10 features results in a dramatic drop in accuracy from approximately 0.57 to 0.31. This precipitous decline suggests that task-relevant information is concentrated into a narrow subset of SAE latents at these stages. We define this point of maximal accuracy drop as the **turning layer**—the layer at which cultural information is most causally concentrated in the processing pipeline, forming a **semantic bottleneck** for cultural knowledge processing. As we demonstrate in Sec. 7, this layer serves as the primary target for subsequent activation steering intervention.

The performance degradation shows a saturation effect regarding the number of features ablated. While there is a noticeable gap between the top-1 and top-10 curves. This indicates that the causal influence is highly localized within a very sparse set of primary latents; once these core features are removed, the remaining latents in the SAE basis contribute marginally to the task, even when a larger number of high-activating features are targeted.

Crucially, ablating an equal number of randomly selected features yields only marginal performance degradation across all layers. This stark contrast establishes that the observed accuracy cliff is not a generic consequence of representation disruption, but instead reflects the targeted removal of semantically privileged latents.

Interestingly, the model exhibits high resilience in the initial (Layers 0–10) and final (Layers 35–40) stages. The stability in early layers suggests features at these depths primarily encode lower-level properties, while the recovery in final layers points

to a stabilized representation where the downstream decision has already been consolidated within the residual stream.

7 Steering for improving Cross Cultural Questions

Having established that cultural information is encoded in a compact set of internal features, in this section we proceed to examine how these representations can be exploited in practice. Concretely, we use culture-specific features as control signals to derive steering vectors, allowing us to systematically influence the model's cultural outputs.

7.1 Experimental Setup

Following the steering framework defined in Section 3, we identify the top-k cultural features based on the joint scores of ν_s^C and Δ_s^C. The intervention is localized to the turning layer(layer 25 for gemma-2-9B and layer 4 for llama-3.1-8B), which are identified in Section 6 as the critical semantic bottleneck for cultural knowledge. To characterize the intensity-response relationship, we systematically vary the steering factor $s \in [0.5, 5]$. Our experiments encompass eight cultures across low(Assam, Northern Nigeria)-, mid-(Greece, Indonesia), and high (China, Iran, Spain, US)-resource settings, following the protocol of Yu et al. (2025a) to ensure generalizability. We benchmark our method against several baselines, including zero-shot prompting, cultural prompting (Tao et al., 2024), and Layer-wise Activation Steering(Turner et al., 2024).

7.2 Contrastive Feature Steering

Based on the steering framework defined in Section 3, we select the top-k cultural features identified by the joint scoring of ν_s^C and Δ_s^C. To refine the steering signal, we implement a weighted contrastive strategy that enhances target cultural attributes while suppressing confounding features. Formally, let I_c and I_w denote the indices of the identified correct and distractor features, respectively. We derive a composite steering direction $\mathbf{v}_s$ from the SAE decoder weights W_{dec} as follows:

$$\mathbf{v}_s = \lambda \cdot \bar{W}_{\text{dec}}^{I_c} - \bar{W}_{\text{dec}}^{I_w} \tag{11}$$

where $\bar{W}_{\text{dec}}^{I}$ represents the mean decoder vector for the corresponding feature set and $\lambda = 2$.

Our intervention is localized to Layers 25, which Section 6 identifies as the pivotal semantic bottleneck for cultural knowledge. During the forward pass, we modulate the intermediate representation h by applying the steering vector: $\Phi(h) = h + s \cdot \mathbf{v}_s$, where $s \in [1, 10]$ is the steering factor. To ensure the generalizability of our findings, we conduct experiments across eight cultures encompassing low-, mid-, and high-resource settings (Yu et al., 2025a). We benchmark our method against zero-shot prompting, cultural prompting (Tao et al., 2024) and Layer-wise Activation Steering(Turner et al., 2024).

7.3 Results

Table 2 shows that SAE Steering consistently achieves the highest average accuracy across both Llama-3.1-8B and Gemma-2-9B. In contrast, both Non-Culture Steering and layer-level activation steering yield only marginal improvements over the zero-shot baseline, indicating that generic or non-targeted interventions on internal representations are insufficient for capturing culturally grounded behaviors. The gains from SAE Steering are particularly pronounced in low-resource cultural settings, where Culture Prompting exhibits unstable or marginal improvements. While Culture Prompting remains competitive in high-resource cultures (e.g., Spain and the US), its performance varies substantially across cultural groups, whereas SAE Steering achieves more balanced improvements and reduces variance between high- and low-resource contexts.

This robustness arises from directly intervening on sparse latent features associated with cultural semantics, rather than relying on surface-level textual cues. To further examine the relationship between these two approaches, we evaluate a hybrid setting that combines Culture Prompting with SAE Steering. However, the hybrid does not yield consistent gains over SAE Steering alone. This suggests that the cultural signal introduced via prompting largely overlaps with the representations captured by the identified SAE features. Once the model is explicitly steered along these latent cultural directions, additional textual cues provide limited or no complementary benefit. Notably, identifying effective steering directions requires only 50 annotated examples per culture, making the approach substantially more data-efficient than fine-tuning-based adaptation. This property is especially relevant for extending foundation models to underrepresented or marginalized cultural contexts, where large-scale supervision is often infeasible.

Although learning SAE representations introduces an upfront computational cost, our findings

Model	Method	AS	CN	ES	GR	ID	IR	NG	US	Avg. Acc
Llama-3.1-8B	Base (Zero-shot)	0.6	0.57	0.63	0.54	0.62	0.59	0.59	0.68	0.58
	Culture Prompting	0.57	0.61	**0.75**	**0.62**	0.66	**0.64**	0.56	**0.76**	0.63
	Non-Culture Steering	0.6	0.58	0.65	0.55	0.63	0.60	0.55	0.68	0.60
	Base + Layer level Steering	0.59	0.61	0.65	0.58	0.60	0.60	0.58	0.72	0.56
	Cultural prompting + SAE Steering	0.63	0.62	0.71	0.61	**0.68**	0.60	0.61	0.74	0.65
	Base + SAE Steering	**0.62**	**0.65**	0.68	0.60	0.67	0.63	**0.62**	0.73	**0.66**
Gemma-2-9B	Base (Zero-shot)	0.49	0.51	0.57	0.50	0.49	0.54	0.49	0.68	0.56
	Culture Prompting	0.57	0.50	0.61	**0.60**	0.59	0.64	0.49	0.74	0.60
	Non-Culture Steering	0.51	0.52	0.58	0.53	0.52	0.58	0.50	0.70	0.55
	Base + Layer level Steering	0.52	0.50	0.60	0.55	0.54	0.60	0.48	0.72	0.58
	Cultural prompting + SAE Steering	0.56	**0.60**	**0.65**	0.58	0.60	**0.70**	**0.53**	0.75	0.62
	Base + SAE Steering	**0.58**	0.58	**0.64**	0.57	**0.60**	0.68	0.52	**0.76**	**0.64**

Table 2: Cross-cultural performance comparison (Accuracy ↑) across eight cultures. We benchmark our contrastive **SAE Steering** against Zero-shot, Culture Prompting and Layer-wise Activation Steering.

suggest a promising paradigm for scalable and equitable cross-cultural adaptation: enabling models to express latent cultural knowledge with minimal supervision, rather than requiring extensive retraining or culturally exhaustive datasets.

8 Conclusion

We present a mechanistic investigation of cultural representations in Large Language Models using Sparse Autoencoders (SAEs), advancing beyond neuron-level approaches that suffer from polysemanticity and superposition. By decomposing intermediate activations into sparse, interpretable feature directions, we identify culture-selective features that remain invariant across paraphrased inputs and task formats—demonstrating that LLMs encode cultural knowledge as abstract semantic structures rather than surface-level lexical patterns.

Through layer-wise causal ablation, we establish that these features are causally necessary for cultural reasoning: removing only the top-10 features at the semantic bottleneck (Layers 22–26) collapses accuracy from 0.57 to 0.31, while matched random ablation causes negligible degradation. Furthermore, activation steering along these feature directions is causally sufficient to modulate culturally grounded outputs across diverse cultural settings without any retraining, achieving consistent improvements over zero-shot and culture-prompting baselines—particularly in low-resource cultural contexts where prompting-based methods are unstable.

Moving beyond the notion of isolated cultural neurons, our work provides a fine-grained and interpretable account of how cultural knowledge is organized within LLMs. This perspective enables more precise analysis and control of culturally grounded behaviors, and offers a principled pathway toward mitigating Western-centric bias and fostering more globally inclusive and culturally adaptive AI systems. We hope this mechanistic framework inspires future work on interpretability-guided cultural alignment across broader model families and languages.

Acknowledgement

Lijie Hu and Chenye Zou are supported by the funding BF0100 from Mohamed bin Zayed University of Artificial Intelligence (MBZUAI).

Limitations

While this study advances the mechanistic understanding of cultural representations in LLMs through Sparse Autoencoders (SAEs), several limitations should be noted. First, our evaluation is primarily based on English-language corpora, which may result in identified features reflecting Western-centric interpretations of culture rather than native nuances. Second, our reliance on social-value surveys (e.g., WVS and GAS) as cultural proxies is inherently reductionist, as these discrete variables may overlook the dynamic, multi-dimensional, and context-dependent nature of cultural identity. Third, SAEs themselves introduce methodological limitations: reconstruction error from the encoder-decoder bottleneck may cause some cultural information to be lost or misattributed, and the phenomenon of feature splitting means that a single cultural concept may be decomposed into multiple fine-grained features across different SAE widths, complicating interpretation. Fourth, the steering factor s in our activation steering experiments re-

quires empirical tuning per culture, and overly large values can degrade generation fluency. A principled method for automatically selecting s remains an open problem. Finally, due to computational constraints, our findings are restricted to specific model families and scales, and their generalizability to larger parameter regimes or non-dense architectures remains to be empirically validated.

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.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Keyuan Cheng, Zijian Kan, Zhuoran Zhang, Muhammad Asif Ali, Lijie Hu, and Di Wang. 2025a. Compke: Complex question answering under knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2557–2576.

Keyuan Cheng, Gang Lin, Haoyang Fei, Lu Yu, Muhammad Asif Ali, Lijie Hu, Di Wang, et al. 2024. Multi-hop question answering under temporal knowledge editing. *arXiv preprint arXiv:2404.00492*.

Keyuan Cheng, Xudong Shen, Yihao Yang, Tengyue-Wang TengyueWang, Yang Cao, Muhammad Asif Ali, Hanbin Wang, Lijie Hu, and Di Wang. 2025b. Codemenv: Benchmarking large language models on code migration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2719–2744.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. *Preprint*, arXiv:2411.14257.

Isabel O. Gallegos, Ryan Aponte, Ryan A. Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, Franck Dernoncourt, Nedim Lipka, Deonna Owens, and Jiuxiang Gu. 2025. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 873–888, Albuquerque, New Mexico. Association for Computational Linguistics.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Arya-

man Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Preprint*, arXiv:2301.04709.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *Preprint*, arXiv:2410.20526.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Representation Learning*, volume 2024, pages 7827–7845.

Xinyan Jiang, Lin Zhang, Jiayi Zhang, Qingsong Yang, Guimin Hu, Di Wang, and Lijie Hu. 2025. Msrs: Adaptive multi-subspace representation steering for attribute alignment in large language models. *arXiv preprint arXiv:2508.10599*.

Difan Jiao, Di Wang, and Lijie Hu. 2026. Understanding the dynamics of demonstration conflict in in-context learning. *arXiv preprint arXiv:2603.04464*.

Hongji Li, Manjiang Yu, Priyanka Singh, Xue Li, Di Wang, Lijie Hu, et al. 2025a. Towards reasoning-preserving unlearning in multimodal large language models. *arXiv preprint arXiv:2512.17911*.

Tong Li, Shu Yang, Junchao Wu, Jiyao Wei, Lijie Hu, Mengdi Li, Derek F. Wong, Joshua R. Oltmanns, and Di Wang. 2025b. Can large language models identify implicit suicidal ideation? an empirical evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18392–18413, Suzhou, China. Association for Computational Linguistics.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *Preprint*, arXiv:2408.05147.

Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. 2024. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Representation Learning*, volume 2024, pages 26486–26515.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass editing memory in a transformer. *The Eleventh International Conference on Learning Representations (ICLR)*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2025. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Christopher Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits.

Pew Global Attitudes Survey. 2022. Pew global attitudes survey. https://www.pewresearch.org/.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Wonduk Seo, Zonghao Yuan, and Yi Bu. 2025. Valuesrag: Enhancing cultural alignment through retrieval-augmented contextual learning. *Preprint*, arXiv:2501.01031.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yi Su, Jiayi Zhang, Shu Yang, Xinhai Wang, Lijie Hu, and Di Wang. 2025. Understanding how value neurons shape the generation of specified values in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9433–9452, Suzhou, China. Association for Computational Linguistics.

Wenjie Sun, Di Wang, and Lijie Hu. 2026. The price of amortized inference in sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*.

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

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *Preprint*, arXiv:2211.00593.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Xinhai Wang, Shu Yang, Liangyu Wang, Lin Zhang, Huanyi Xie, Lijie Hu, and Di Wang. 2025. Pahq: Accelerating automated circuit discovery through mixed-precision inference optimization. *arXiv preprint arXiv:2510.23264*.

World Values Survey. 2022. World values survey. `https://www.worldvaluessurvey.org/wvs.jsp`.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Lijie Hu, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, et al. 2024. Usable xai: 10 strategies towards exploiting explainability in the llm era. *arXiv preprint arXiv:2403.08946*.

Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. 2025. Neuron-level analysis of cultural understanding in large language models. *Preprint*, arXiv:2510.08284.

Shu Yang, Shenzhe Zhu, Zeyu Wu, Keyu Wang, Junchi Yao, Junchao Wu, Lijie Hu, Mengdi Li, Derek F Wong, and Di Wang. 2025. Fraud-r1: A multi-round benchmark for assessing the robustness of llm against augmented fraud and phishing inducements. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4374–4420.

Junchi Yao, Shu Yang, Jianhua Xu, Lijie Hu, Mengdi Li, and Di Wang. 2025. Understanding the repeat curse

in large language models from a feature perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7787–7815.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *EMNLP*.

Haeun Yu, Seogyeong Jeong, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and Isabelle Augenstein. 2025a. Entangled in representations: Mechanistic investigation of cultural biases in large language models. *Preprint*, arXiv:2508.08879.

Manjiang Yu, Hongji Li, Priyanka Singh, Xue Li, Di Wang, and Lijie Hu. 2025b. Pixel: Adaptive steering via position-wise injection with exact estimated levels under subspace calibration. *arXiv preprint arXiv:2510.10205*.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *ICLR*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2025a. Evaluating and improving cultural awareness of reward models for llm alignment. *Preprint*, arXiv:2509.21798.

Lin Zhang, Wenshuo Dong, Zhuoran Zhang, Shu Yang, Lijie Hu, Ninghao Liu, Pan Zhou, and Di Wang. Eap-gp: Mitigating saturation effect in gradient-based automated circuit identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lin Zhang, Lijie Hu, and Di Wang. 2025b. Mechanistic unveiling of transformer circuits: Self-influence as a key to model reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1387–1404.

Zhuoran Zhang, Yongxiang Li, Zijian Kan, Keyuan Cheng, Lijie Hu, and Di Wang. 2025c. Locate-then-edit for multi-hop factual recall under knowledge editing. In *International Conference on Machine Learning*, pages 75369–75391. PMLR.

Appendix

A Prompt Used for Rewrite Questions

Here is the prompt we used to rewrite the questions.

```
You are given an input query written
as a question.
Rewrite this query into 5 alternative sentences
that aresemantically equivalent
to the original query.

Requirements:
- Preserve the original intent and meaning.
- Maximize linguistic and structural diversity
across the 5 rewrites.
- The rewrites do not need to remain questions;
you may convert them into cloze-style statements,
declarative sentences, conditional forms,
or other appropriate formats.
- Vary phrasing, syntax, and
perspective where possible.
- Do not introduce new i
nformation or assumptions.

Output only the 5 rewritten
sentences, each on a separate line.
```

B Experiment Settings

For each experiment, we don't perform any sampling during generation to avoid randomness. For SAEs from Gemma Scope(Lieberum et al., 2024), we choose 16k width and the one with the first smallest L0 value for each layer. For SAEs from Llama Scope(He et al., 2024), we use the model from `https://huggingface.co/OpenMOSS-Team/Llama3_1-8B-Base-LXR-8x`.

C Jaccard Similarity across different selected features

Additional results for Jaccard Similarity across different selected features for two different LLMs are demonstrated in Figure 5 and 6.As we select more features, the similarity begins to decrease because most feature values are below 1, leading to lower similarity and thus making the distinction between the question and its closed form increasingly insignificant.

D Additional Results for Culture-injection Effects

Additional results for Culture-injection Effects of the value of Δ_s^C across different layers for two different LLMs are demonstrated in Figure 7 and 8. Almost all layers also show similar conclusions to Figure 2.

E Additional Results for Culture-specificity scores

Additional results for Culture-specificity scores of the value of v_s^C across different layers for two different LLMs are demonstrated in Figure 9 and 10.

Figure 5: Layer-wise Jaccard similarity of selected top-k culture-selective SAE features between Question (QA) and Cloze formats on Llama-3.1-8B.

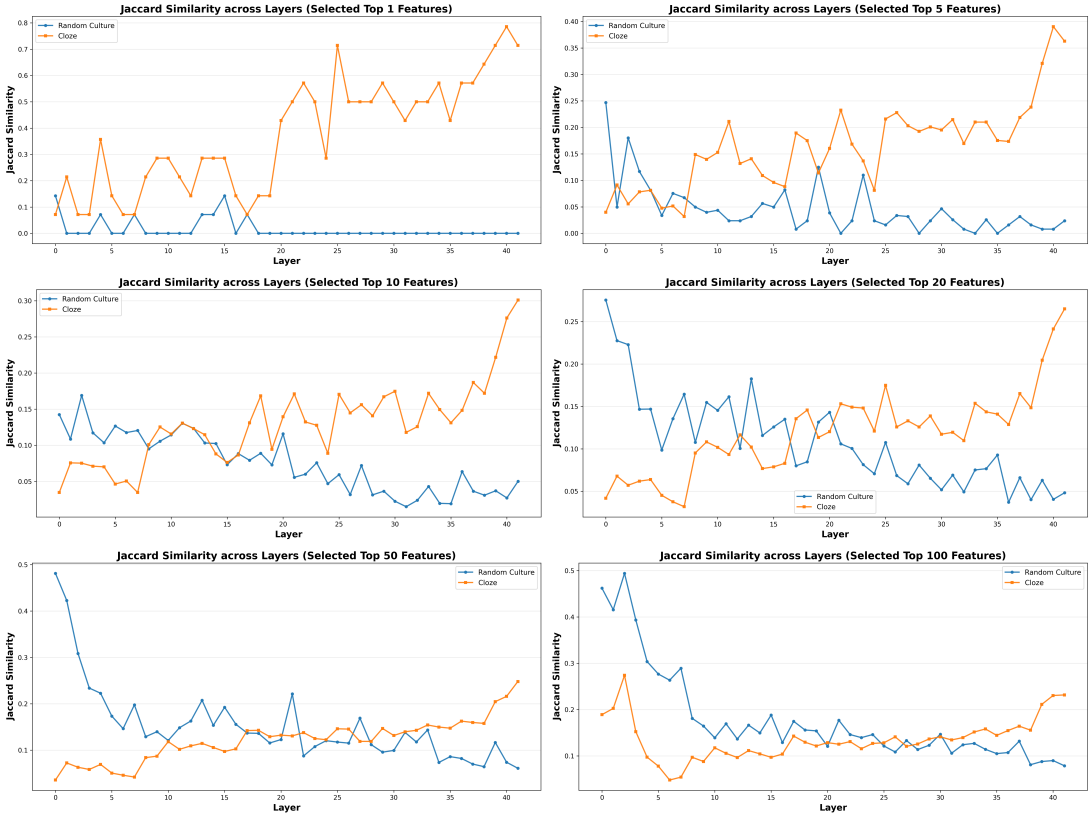


Figure 6: Layer-wise Jaccard similarity of selected top-k culture-selective SAE features between Question (QA) and Cloze formats on gemma-2-9B.

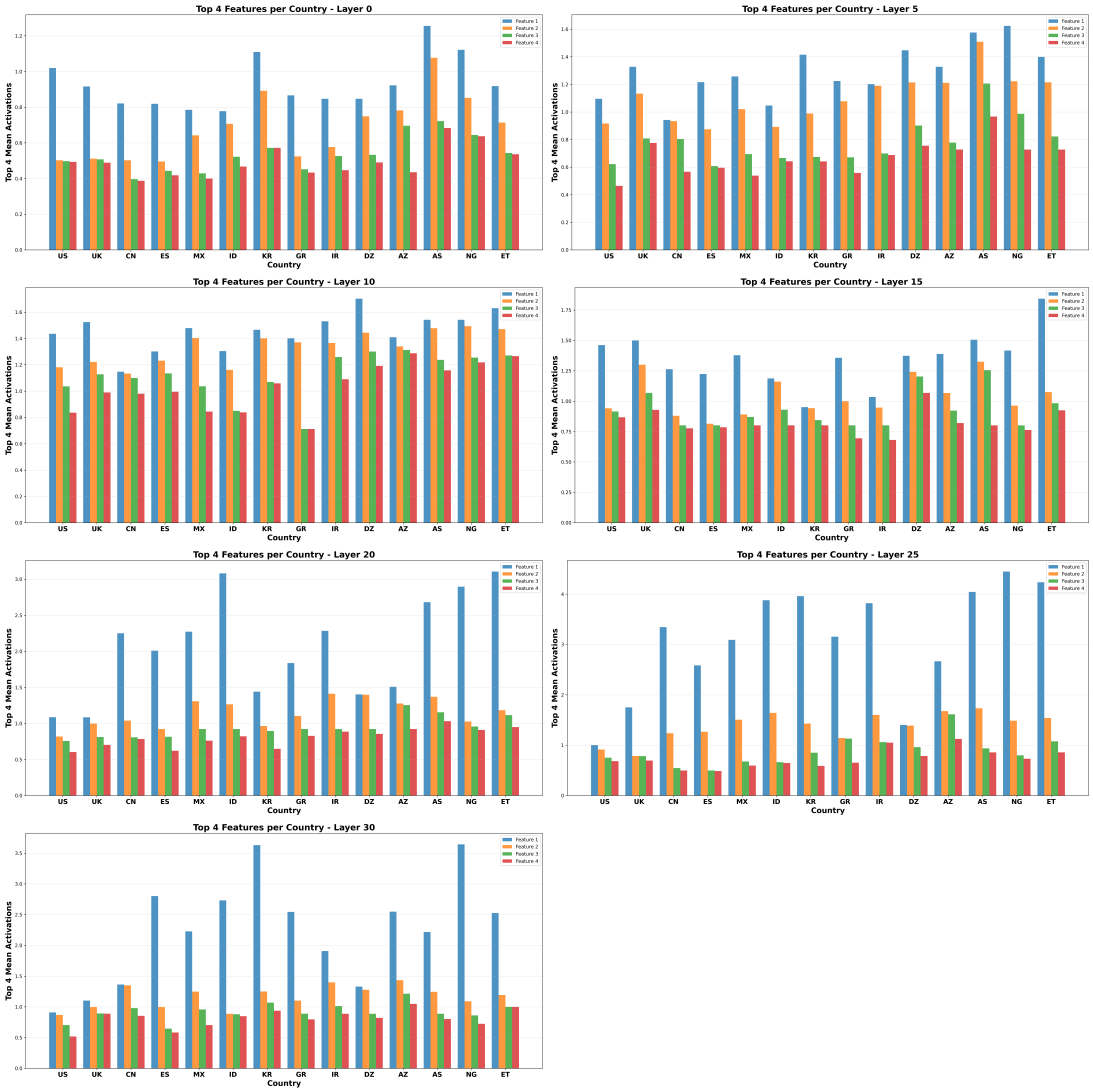


Figure 7: The value of Δ_s^C of Llama-3.1-8B across different layers

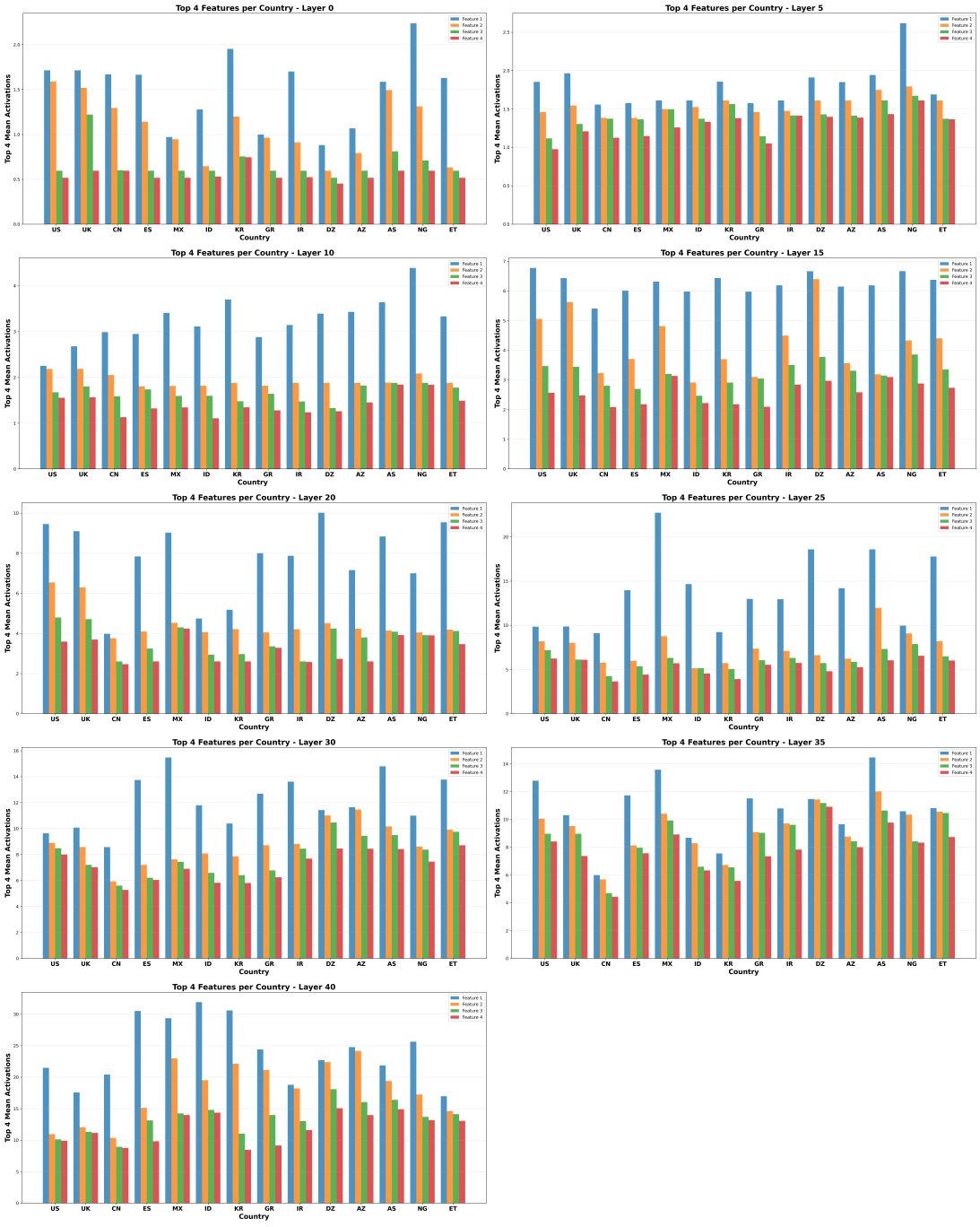


Figure 8: The value of Δ_s^C of Gemma-2-9B across different layers

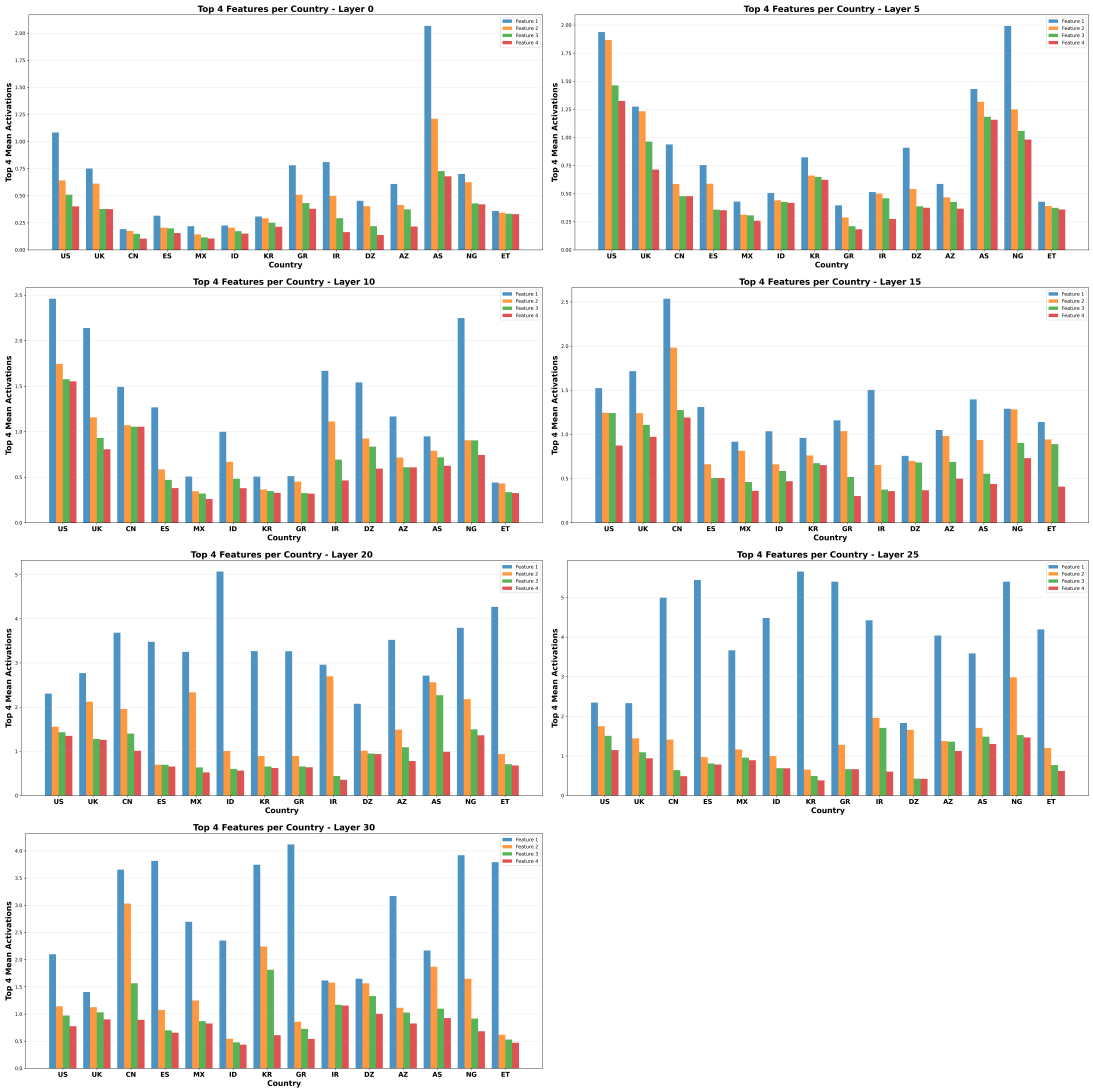


Figure 9: The value of v_s^C of Llama-3.1-8B across different layers

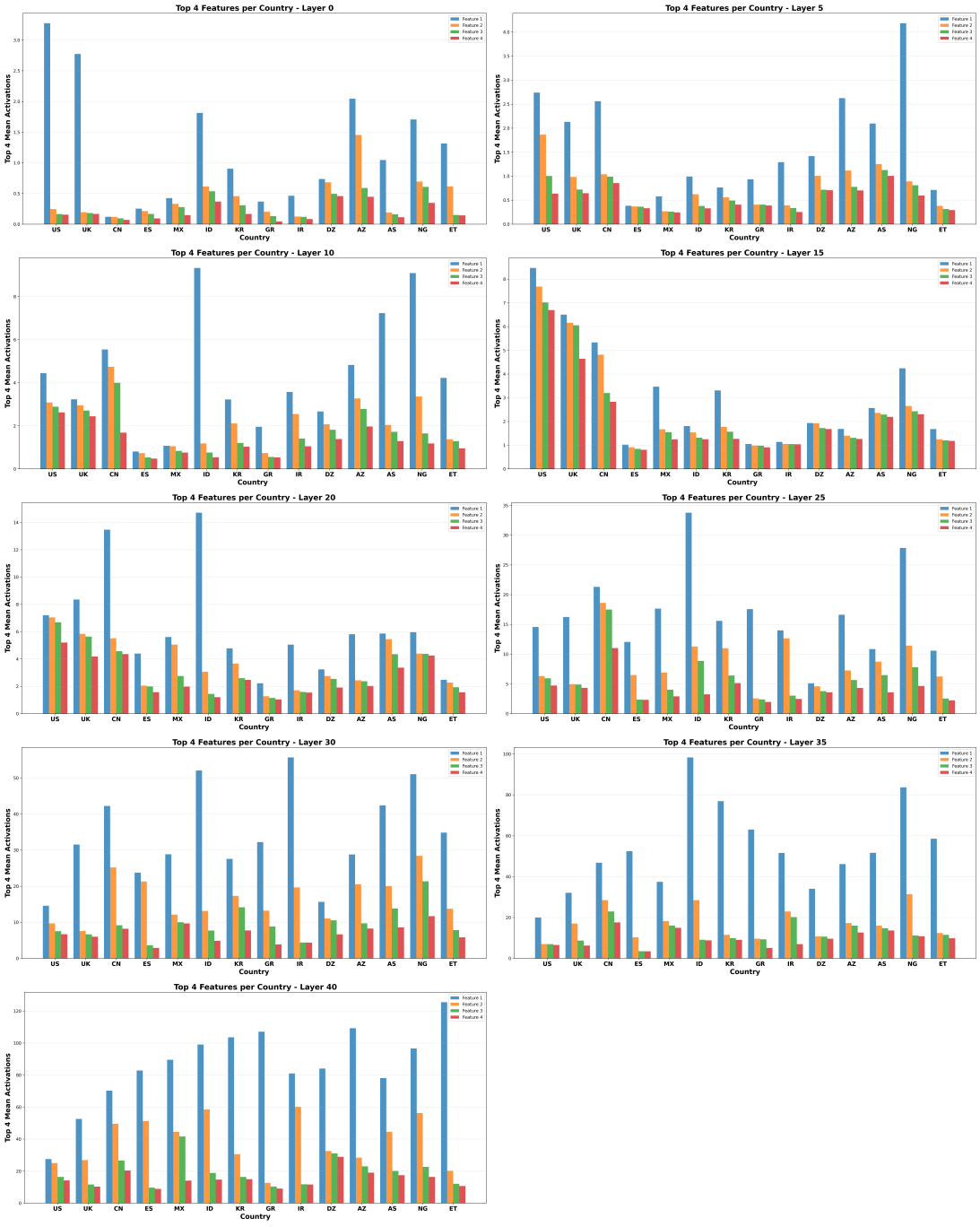


Figure 10: The value of v_s^C of Gemma-2-9B across different layers

F Additional Results for Directional Ablation

Additional results for directional ablation of different cultures across different layers. And almost every country shows similar turning point at all most the same layer.

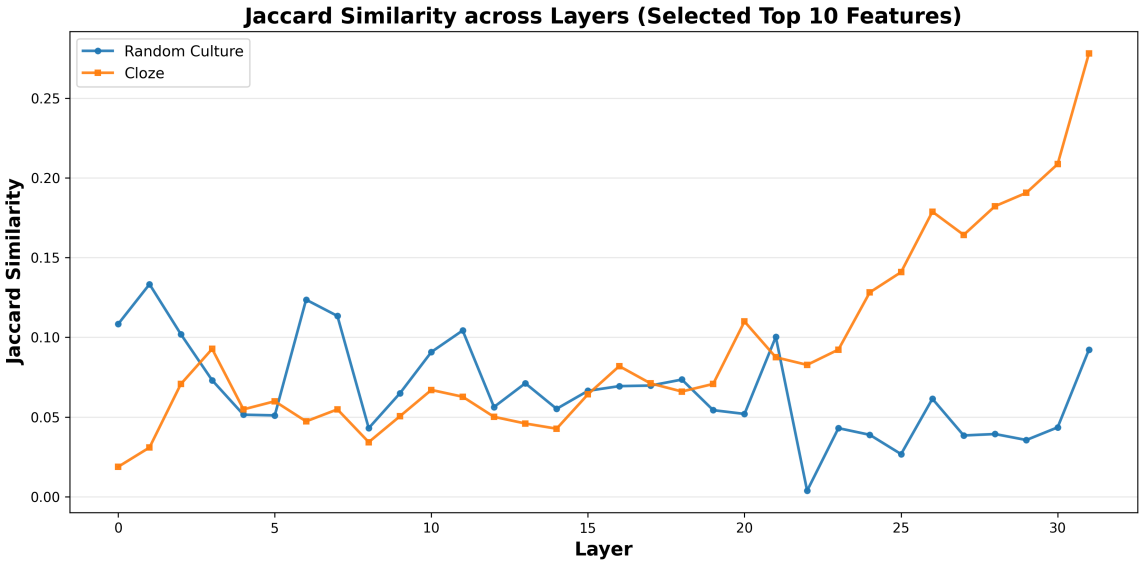


Figure 11: Layer-wise accuracy under top-k SAE feature ablation of Llama-3.1-8B. Random feature ablation with matched cardinality is shown as a negative control.

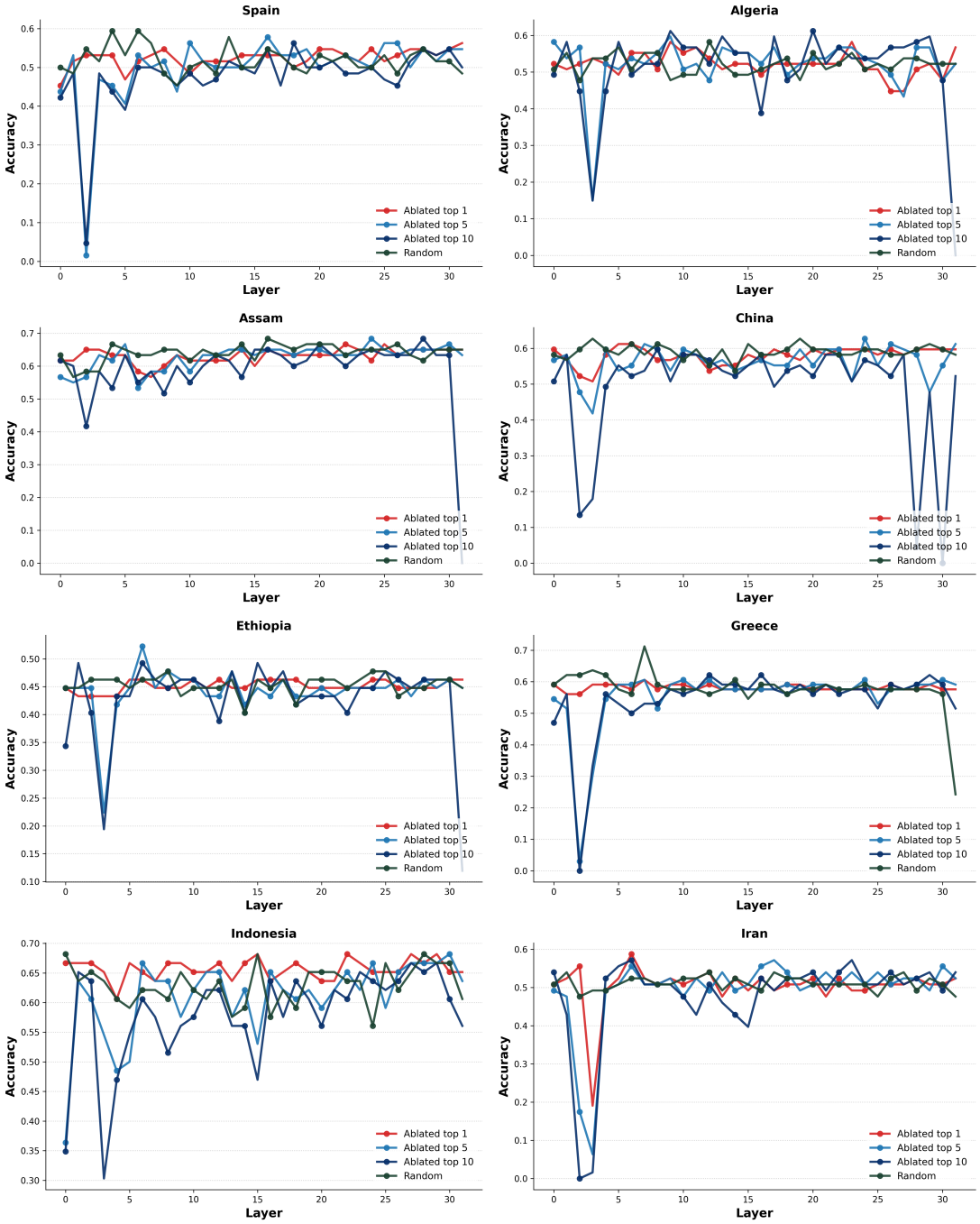


Figure 12: Layer-wise accuracy under top-k SAE feature ablation across different countries of Llama-3.1-8B.

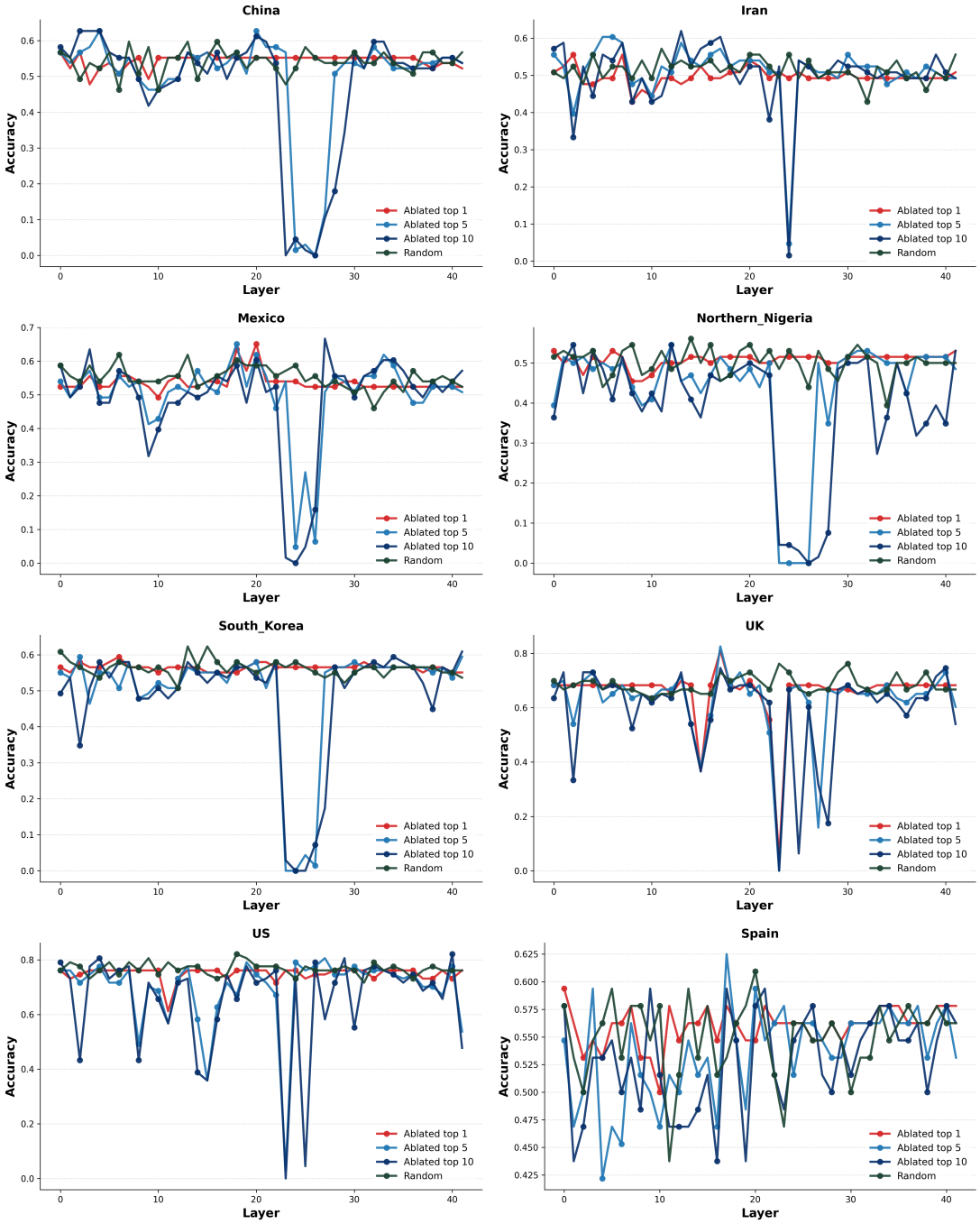


Figure 13: Layer-wise accuracy under top-k SAE feature ablation across different countries of Gemma2-9B.