Isolating Culture Neurons in Multilingual Large Language Models

Danial Namazifard
University of Tehran
namazifard@ut.ac.ir

Lukas Galke Poech
University of Southern Denmark
galke@imada.sdu.dk

Abstract

Language and culture are deeply intertwined, yet it has been unclear how and where multilingual large language models encode culture. Here, we build on an established methodology for identifying language-specific neurons to localize and isolate culture-specific neurons, carefully disentangling their overlap and interaction with language-specific neurons. To facilitate our experiments, we introduce MUREL, a curated dataset of 85.2 million tokens spanning six different cultures. Our localization and intervention experiments show that LLMs encode different cultures in distinct neuron populations, predominantly in upper layers, and that these culture neurons can be modulated largely independently of language-specific neurons or those specific to other cultures. These findings suggest that cultural knowledge and propensities in multilingual language models can be selectively isolated and edited, with implications for fairness, inclusivity, and alignment. Code and data are available at https://github.com/namazifard/Culture_Neurons.

1 Introduction

Cultural context underpins human communication, shaping interpretations, values, and worldviews that go beyond linguistic surface forms (Kramsch, 2014). For example, opinions on morality, authority, and gender roles can vary dramatically across cultural groups, even when expressed in the same language. Recent advances in multilingual large language models (LLMs) (Team et al., 2025) have drawn increased attention on their cultural propensities. Understanding and potentially controlling the cultural propensities of language models is crucial to ensure cultural fairness, inclusivity, and alignment (Liu et al., 2025).

Here, we set out to study model internals governing such cultural propensities. Can we localize a set of neurons that drives cultural propensities? Is this set of neurons separate from the language associated with that culture? Can we intervene on the model internals to modulate cultural propensities without any training?

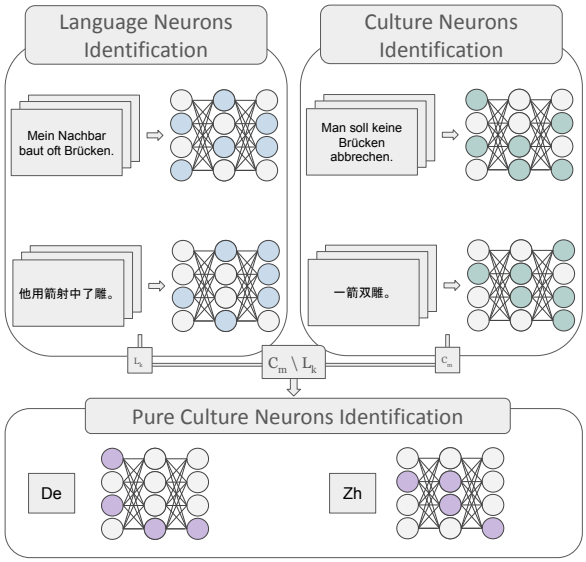


Figure 1: Overview of our methodology for identifying pure culture-specific neurons in language models. We first identify language-specific neurons ($\mathbb{L}_k$) using literal, language-focused sentences (left), and culture-specific neurons ($\mathbb{C}_m$) using culturally salient phrases (right). By subtracting the language-specific neuron set from the culture-specific neuron set, we obtain pure culture-specific neurons ($\mathbb{C}_m \setminus \mathbb{L}_k$), which encode culture independently of language (bottom).

Despite advances in neuron localization and editing techniques (Dai et al., 2022; Hou et al., 2023; Li et al., 2023a), isolating *pure culture-specific neurons*, i.e., those that drive culture but not language, remains particularly challenging, given their inherent entanglement (Liu et al., 2025; Kramsch, 2014) and even data on non-linguistic elements of culture needs to be encoded linguistically when fed into a language model.

Prior work on localization of language-specific neurons has revealed that different languages are

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 768–785

encoded in different areas of the model (Tang et al., 2024; Zhao et al., 2024). However, these methods are insufficient for identifying and isolating culture-specific neurons, due to the expected entanglement with language and the lack of suitable datasets.

To address these challenges, we develop a methodology for identifying both culture-specific and pure culture-specific neurons, by which we mean those neurons that are specific to a culture but not specific to the language associated with that culture. To facilitate this methodology, we compile a dataset covering different linguistic and non-linguistic cultural elements.

Our results indicate that, despite the inherent entanglement of language and culture, it is possible to identify neuron populations more strongly associated with culture than language. We find that cultural representations are localized in regions often distinct from language-encoding neurons, with different cultures occupying separate neural populations. This allows selective modulation of one culture's representations largely independently of other cultures.

In sum, the contributions of this work are:

- We introduce a methodology for identifying (pure) culture neurons in LLMs (§3).
- We present MUREL, a dataset of 85.2 million tokens covering six cultures (§4).
- We conduct experiments showing that culture-specific neurons are largely separable from language-specific neurons (§5 and 6.1).
- We show that culture neurons can be selectively modulated (§6.2).

2 Related Work

Recent work on multilingual language models has primarily focused on linguistic capabilities (Liang et al., 2022; Srivastava et al., 2023; Ahuja et al., 2023). More recently, research has shifted toward assessing cultural competence, including culturally salient elements such as norms (Ziems et al., 2023), values (Moore et al., 2024), and worldviews (Mushtaq et al., 2025). Some studies have examined cultural knowledge within a monocultural setting (Müller-Eberstein et al., 2025), while a growing body of work investigates multicultural evaluations, exploring cultural phenomena across languages and societies (Yin et al., 2022; Fung et al., 2023; Huang et al., 2025). Recent efforts have extended this research to vision-language models, evaluating cultural understanding within monocultural and multicultural contexts (Alwajih et al., 2024; Romero et al., 2024; Vayani et al., 2025; Nayak et al., 2024).

Despite these advances, existing evaluations are predominantly behavioral, leaving open critical questions about how cultural knowledge is internally encoded in multilingual models.

In parallel, mechanistic interpretability research has sought to uncover how LLMs encode information at the neuron level. These studies have successfully identified neurons or populations of neurons corresponding to specific capabilities, such as knowledge storage (Dai et al., 2022), safety alignment (Chen et al., 2024), and confidence estimation (Stolfo et al., 2024). Recent work has further explored specialized neurons in multilingual models, discovering neurons encoding language identity or linguistic features (Tang et al., 2024; Zhao et al., 2024; Kojima et al., 2024), with similar findings for vision-language models (Huo et al., 2024).

However, while these advances have deepened our understanding of language models' representations of linguistic and semantic information, they have not addressed the neural encoding of cultural knowledge independent of linguistic identity. This paper addresses this gap by introducing a systematic methodology for studying culture-specific neurons, providing new insights into the cultural representations in language models.

3 Identifying Culture Neurons

Our goal is to identify and isolate *culture-specific neurons* within multilingual LLMs. To disentangle language and culture, we develop a systematic approach to distinguish neuron populations that respond specifically to cultural inputs, independently of linguistic features. This requires first identifying language-specific and culture-specific neurons, and then applying set operations to isolate pure culture neurons, as described below.

3.1 Background: Identification of Language-Specific Neurons

We first locate *language-specific neurons* as the basis for disentangling linguistic and cultural factors. We adopt the language activation probability entropy (LAPE) method (Tang et al., 2024), which effectively detects language-localized regions in multilingual LLMs. We briefly recapitulate their approach, as it forms the foundation for our identification of culture-specific neurons.

Modern LLMs are built on autoregressive trans-

former architectures (Vaswani et al., 2017) with multi-head self-attention (MHA) and feed-forward networks (FFNs). Let $\tilde{\boldsymbol{h}}^{\ell}$ denote the output of the MHA module in the ℓ-th layer, computed using the previous layer's hidden states and trainable parameters, and $\mathrm{act\_fn}(\cdot)$ denotes the activation function. The FFN output $\boldsymbol{h}^{\ell} \in \mathbb{R}^{d_1}$ in a GLU variant is:

$$\boldsymbol{h}^{\ell} = (\mathrm{act\_fn}(\tilde{\boldsymbol{h}}^{\ell}\mathbf{W}_1^{(\ell)}) \otimes \tilde{\boldsymbol{h}}^{\ell}\mathbf{W}_3^{(\ell)}) \cdot \mathbf{W}_2^{(\ell)},$$

where $\mathbf{W}_1^{(\ell)}, \mathbf{W}_3^{(\ell)} \in \mathbb{R}^{d_1 \times d_2}$ and $\mathbf{W}_2^{(\ell)} \in \mathbb{R}^{d_2 \times d_1}$ are learnable parameters. In LAPE, a *neuron* is defined as the linear transformation of a single column in $\mathbf{W}_1^{(\ell)}$ followed by the application of the non-linear activation function. Thus, each FFN module contains d_2 neurons.

LAPE identifies neurons with systematically different activation probabilities across languages. For neuron j in layer ℓ and language k:

$$p_{\ell,j}^{k} = \mathbb{E}\left[\mathbb{I}(\mathrm{act\_fn}(\tilde{\boldsymbol{h}}^{\ell}\mathbf{W}_1^{(\ell)})_j > 0) \,\middle|\, \text{language } k\right],$$

where $\mathbb{I}(\cdot)$ is the indicator function. The activation probability is empirically estimated by the likelihood that the neuron's activation value exceeds zero. The probabilities across all languages $\mathcal{L}$ yield a distribution $\boldsymbol{p}_{\ell,j} = (p_{\ell,j}^{1}, ..., p_{\ell,j}^{k}, ..., p_{\ell,j}^{l})$, which is normalized: $p'^{k}_{\ell,j} = \frac{p_{\ell,j}^{k}}{\sum_{k' \in \mathcal{L}} p_{\ell,j}^{k'}}$ The entropy of this distribution is: $\mathrm{LAPE}_{\ell,j} = -\sum_{k \in \mathcal{L}} p'^{k}_{\ell,j} \log p'^{k}_{\ell,j}$ Neurons with low LAPE are highly language-specific. We define language-specific neurons as those in the bottom 1% of LAPE, requiring that at least one language has an activation probability above a specified threshold. In practice, we follow Tang et al. (2024) by using balanced corpora per language and computing LAPE scores, yielding a sparse set $\mathbb{L}_k$ for each language k.

3.2 Identification of Culture-Specific Neurons

Similarly, we define *Culture Activation Probability Entropy* (CAPE) by evaluating activation probabilities over culturally distinct inputs. For culture m, the activation probability for neuron j in layer ℓ is defined as:

$$q_{\ell,j}^{m} = \mathbb{E}\left[\mathbb{I}(\mathrm{act\_fn}(\tilde{\boldsymbol{h}}^{\ell}\mathbf{W}_1^{(\ell)})_j > 0) \,\middle|\, \text{culture } m\right],$$

We normalize and compute entropy as with LAPE, and define *culture-specific neurons* $\mathbb{C}_m$ as those with CAPE below the threshold τ_{cult}.

$$\mathbb{C}_m = \{v \in \mathbb{N} \mid \mathrm{CAPE}(v) \leq \tau_{\text{cult}}\}$$

3.3 Disentangling Culture from Language

To disentangle language and culture, we apply set operations at the neuron level. Let $\mathbb{N}$ denote the set of all FFN neurons in a given model. For each language k, we identify a subset $\mathbb{L}_k \subset \mathbb{N}$ of *language-specific neurons* using the LAPE method (§3.1). Similarly, for each culture m, we define the set $\mathbb{C}_m \subset \mathbb{N}$ of *culture-specific neurons* (§3.2).

We assume that there is some overlap between language and culture neurons. To isolate *pure culture-specific neurons* for culture m and its associated language k, we define: $\mathbb{P}_m = \mathbb{C}_m \setminus \mathbb{L}_k$ as the set of neurons specific to culture m that are not language-specific. Some neurons may respond to both language and culture, which we define as *compound language-and-culture neurons* $\mathbb{L}_k \cap \mathbb{C}_m$. This framework partitions neuron space into language, culture, compound, and generic components.

3.4 Interventions

We assess the functional roles of these neuron subsets by systematically deactivating (zeroing out) their activations during inference.

Deactivating neuron subpopulations: Given a set of neurons $\mathbb{X}$ – where $\mathbb{X} \in \{\mathbb{L}_k, \mathbb{C}_m, \mathbb{P}_m, \mathbb{L}_k \cap \mathbb{C}_m\}$, representing language-specific, culture-specific, pure-culture, and compound neurons, respectively – we set all activations in $\mathbb{X}$ to zero.

Random neuron ablation: As a control, we select and deactivate a size-matched amount of random neurons to ensure observed effects are not due to neuron count alone.

4 MUREL: A Multicultural Resource for Evaluating Language Models

To support our neuron analysis, we introduce **MUREL** (**MU**lticultural **R**esource for **E**valuating **L**anguage Models), a comprehensive dataset collection spanning culturally diverse text resources. MUREL is constructed from public sources and systematically organized according to the taxonomy proposed by Liu et al. (2025), enabling broad coverage of ideational, linguistic, and social dimensions for targeted analysis of culture-specific and linguistic phenomena. In total, MUREL comprises 69 datasets spanning six cultural groups, containing an average of 14.2 million tokens per culture (see Appendix A for detailed statistics).

4.1 Dataset Organization

We categorize the datasets into three primary branches, as defined by Liu et al. (2025): (i) **Ideational Elements**, covering abstract cultural concepts and knowledge; (ii) **Linguistic Elements**, focusing on intra-linguistic variations and communicative styles; and (iii) **Social Elements**, encompassing factors related to human interactions and demographic attributes. An overview of the datasets along these dimensions is described below.

Ideational Elements comprise concepts, knowledge, values, norms, morals, and artifacts. Concepts are salient, lexicalized ideas representing either culturally unique objects or figurative expressions, for which we use data on metaphors (Kabra et al., 2023), proverbs and sayings (Liu et al., 2024), idioms (Stap et al., 2024; Khoshtab et al., 2025), and ironies (Casola et al., 2024). Knowledge: Culture-specific factual and common-sense information is covered through cultural probing datasets (Bhatt and Diaz, 2024), multiple-choice QA benchmarks (Wang et al., 2024), and knowledge bases capturing cultural knowledge (Koto et al., 2024).

Values represent beliefs and behavioral standards prioritized differently across cultural groups. To capture these, we combine established resources such as the Pew Global Attitudes Survey (PEW)[1], World Values Survey (WVS)[2], Political Compass Test (PCT)[3], and Hofstede's Cultural Dimensions (Hofstede, 1984). Additionally, we consider recent NLP datasets specifically developed to assess the alignment and manifestation of cultural values in large language models (Cao et al., 2023; Pistilli et al., 2024; Lee et al., 2024a).

Norms and Morals are sets of culture-dependent principles governing acceptable behaviors and judgments. To cover this area, we utilize existing norm banks (Dwivedi et al., 2023; CH-Wang et al., 2023; Fung et al., 2023). Additionally, we incorporate datasets that employ direct querying of language models on ethical and normative issues (Yuan et al., 2024; Yu et al., 2024).

Artifacts include culturally significant products of human creativity such as literature, poetry, music, films, and memes. Our compilation incorporates datasets covering literary texts, fairy tales, and poetry, designed explicitly for cultural analysis and cross-cultural adaptation (Yang et al., 2019; Chakrabarty et al., 2021; Schmidt et al., 2021).

Linguistic Elements cover dialects, styles, registers, and genres. Dialects are systematic linguistic variants influenced by regional, national, or sociocultural factors.

To encompass dialectal diversity, our compilation integrates datasets designed for dialect identification and analysis (Malmasi and Zampieri, 2017; Ciobanu et al., 2018) as well as resources focusing on translations between dialects and standard languages (Plüss et al., 2023; Kuparinen et al., 2023). Styles, Registers, Genres include linguistic variations shaped by situational context, communicative goals, and societal norms. Our compilation incorporates datasets designed to evaluate style and register in NLP tasks, focusing on aspects such as formality (Nadejde et al., 2022), politeness (Srinivasan and Choi, 2022; Havaldar et al., 2023), slang (Sun and Xu, 2022), and genre-specific language, including news reporting and storytelling.

Social Elements cover relationships, context, communicative goals, and demographics. Relationship addresses how communication varies according to interpersonal and societal connections, such as family roles or social hierarchies. Our collection includes datasets that explicitly account for culture-specific relationship terms and interaction dynamics (Zhan et al., 2024), ensuring nuanced modeling of communication styles sensitive to relationship contexts.

Context refers to the linguistic and extra-linguistic settings shaping communication, such as situational, historical, or non-verbal cues. To comprehensively address contextual variation, our dataset compilation includes resources emphasizing both textual contexts and broader frames of reference (Hovy et al., 2020; Chakrabarty et al., 2022a; Zhan et al., 2023; Ziems et al., 2023).

Communicative Goals cover culturally distinct purposes behind language use, including indirect versus direct communication styles in refusal, requests, and apologies. We incorporate resources tailored to evaluating these pragmatic variations, supporting tasks that require understanding shaped communicative intents and their linguistic expressions (Emelin et al., 2021; Li et al., 2023b; Zhan et al., 2024).

Demographics reflect characteristics of individuals and groups, such as age, income, educational level, or ethnicity, which influence commu-

[1] https://www.pewresearch.org/
[2] https://www.worldvaluessurvey.org/
[3] https://www.politicalcompass.org/

nication patterns. Our dataset selection includes demographic-focused datasets that facilitate exploration of how sociodemographic attributes impact linguistic usage and perception (Voigt et al., 2018; Hovy et al., 2020; Santy et al., 2023).

4.2 Language Selection

For our study, we selected six typologically and geographically diverse languages: English (*en*), German (*de*), Danish (*da*), Chinese (*zh*), Russian (*ru*), and Persian (*fa*). The selection was guided by three criteria: (a) **geographical diversity**, covering Western Europe, East Asia, Eastern Europe, and the Middle East; (b) **linguistic typology**, including the Germanic, Slavic, Indo-Iranian, and Sino-Tibetan language families; and (c) **resource availability**, spanning both high-resource (e.g., English) and lower-resource (e.g., Danish) languages. This diversity enables us to assess the robustness of neuron detection methods across a broad spectrum of linguistic and cultural contexts, enhancing the generalizability of our findings beyond any single language, culture, or region.

4.3 Dataset Preparation

To ensure that each dataset was suitable for detecting culture-specific neurons, we performed targeted adaptation and reformatting for several datasets. While some datasets could be directly integrated in their original format, others required modification to better align with our experimental setup and enabling finer-grained cultural analysis.

For example, we transformed original World Values Survey (WVS) probes into textual statements that explicitly encode cultural nuances. Original survey items, such as, "Familie ist [MASK] in meinem Leben" with possible responses "wichtig" or "unwichtig" were converted into complete statements (e.g., *Familie ist wichtig in meinem Leben*). Such transformations allow us to treat each response as an independent cultural assertion and standardize inputs while preserving cultural information.

5 Experimental Setup

Our goal is to systematically identify **pure culture-specific neurons** that respond specifically to cultural information, independent of language, within multilingual LLMs. We proceed as follows:

We first apply the language activation probability entropy (LAPE) method (§3.1) to Wikipedia corpora[4]. For each language k, we use 100 million tokens to robustly capture neuron activation patterns across diverse linguistic contexts, following established methodology (Tang et al., 2024).

Next, we apply the culture activation probability entropy (CAPE) method (§3.2) to our MUREL dataset, using 10 million tokens per culture m. Note that for each culture m, CAPE is computed over the *union* of texts sampled from all three MUREL branches, i.e., the individual branches are *not* considered as separate CAPE targets. For evaluation, we use a separate, balanced held-out set of 100,000 tokens per culture, ensuring reliable measurement of neuron specificity.

Following prior work (Tang et al., 2024), we select the lowest 1% of neurons by entropy as language- and culture-specific neurons for LAPE and CAPE, respectively. We use 1% for sparsity and comparability. Prior studies found stable trends across cutoffs ranging from 1 to 10%, and our pre-experiments were consistent with these findings.

To disentangle the effects of language and culture, we categorize identified neurons as (1) *Pure culture-specific neurons*: neurons that respond strongly to culture m but are not language-specific; and (2) *Compound language-and-culture neurons*, which respond to both language k and culture m.

We conduct our experiments using four transformer-based pretrained language models, including `Llama-2-7b` (Touvron et al., 2023), `Llama-3.1-8b` (Grattafiori et al., 2024), `Qwen2.5-7b` (Yang et al., 2025), and `Gemma-3-12b` (Team et al., 2025).

All models except Llama-2 are multilingual; Llama-2 is included as a monolingual baseline to test how our methodology generalizes beyond multilingual settings. Additional details for each model are provided in Appendix B.

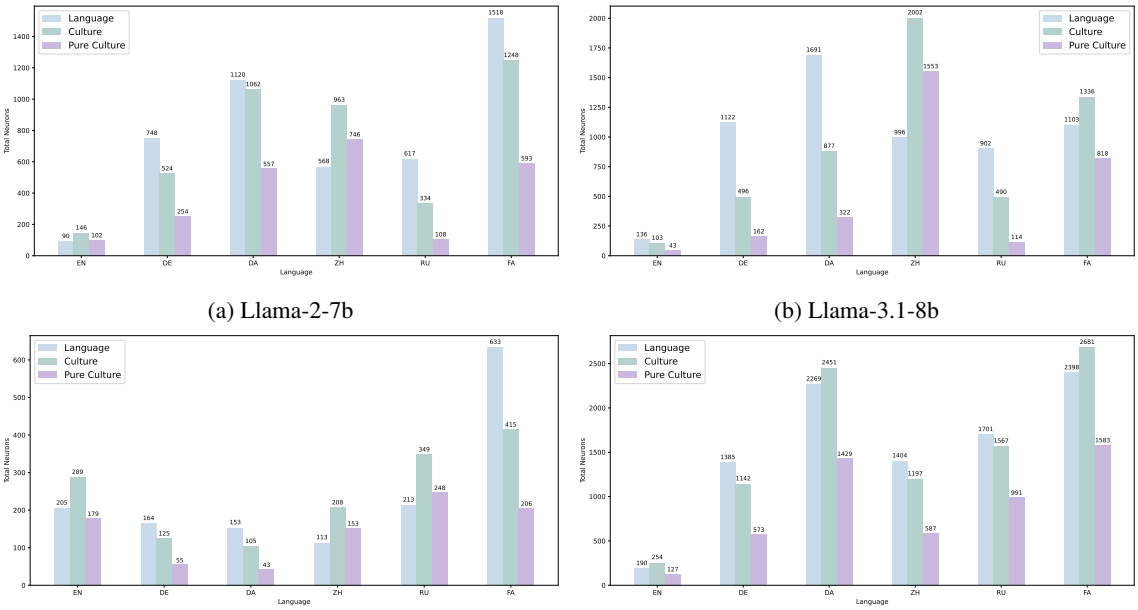
6 Results

We first report neuron identification and distribution, then run intervention experiments to assess functional roles.

6.1 Neuron Identification and Distribution

Neuron Counts and Distributions Language- and culture-specific neurons were selected from all FFN layers based on the lowest activation entropy values. Figure 2 shows the number of language-

[4] https://huggingface.co/datasets/wikimedia/wikipedia



(a) Llama-2-7b

(b) Llama-3.1-8b

(c) Qwen2.5-7b

(d) Gemma-3-12b

Figure 2: Total language, culture, and pure culture neurons per language and model. The total number of identified neurons is 3,523 for Llama-2-7b, 4,588 for Llama-3.1-8b, 1,004 for Qwen2.5-7b, and 7,373 for Gemma-3-12b.

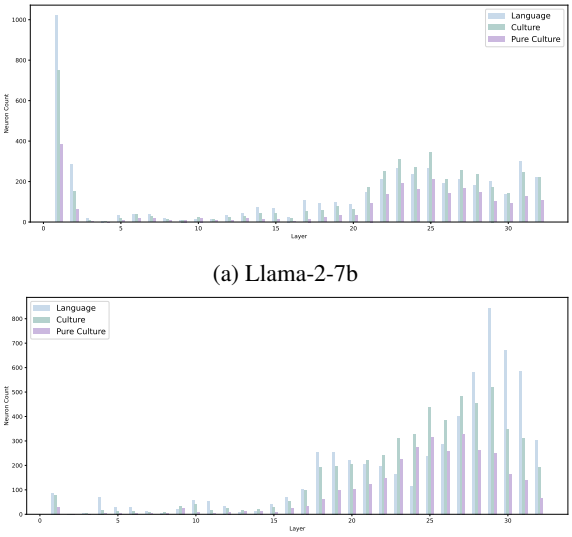
specific, culture-specific, and pure culture-specific neurons identified across the four evaluated models for each tested language and culture.

Several notable patterns emerge: First, the degree of neuron specialization varies not only by model but also across languages and cultures, reflecting distinct representational demands.

Second, lower-resource languages such as Persian and Danish show higher counts of both language-specific and culture-specific neurons compared to resource-rich languages like English. This aligns with prior work (Tang et al., 2024), suggesting that multilingual models allocate more representational capacity to underrepresented languages to capture richer linguistic and cultural nuances.

Third, although we explicitly target exactly 1% of the neurons within FFN layers for both language- and culture-specific sets, we observe a slight discrepancy in the total neuron counts. This minor discrepancy arises naturally because some neurons simultaneously encode multiple languages or cultures, resulting in overlapping neuron sets.

Crucially, a substantial proportion (on average 56.7%) of culture-specific neurons are categorized as *pure culture-specific*, indicating they encode cultural representations largely independent of linguistic identity. This suggests that much of the cultural information within multilingual LLMs is neurally localized to specific populations of neurons that are, to a considerable degree, separable from language processing.

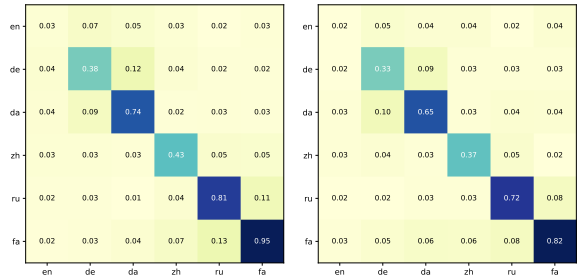


(a) Llama-2-7b

(b) Llama-3.1-8b

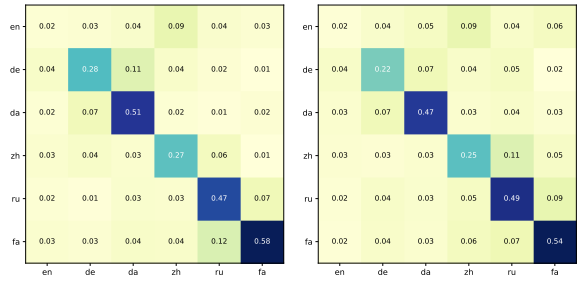
Figure 3: Layer-wise distribution of language-, culture-, and pure culture-specific neurons for models. Layer-wise distribution per language is shown in Figure 13.

Neuron Distribution across Layers. We next examine how neuron types are distributed across model layers. As shown in Figure 3, models tend to concentrate language- and culture-specific neurons in the upper layers. In Llama-2-7b, a monolingual



(a) Culture

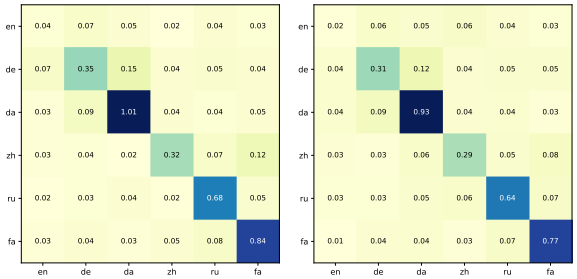
(b) Pure Culture



(c) Language ∩ Culture

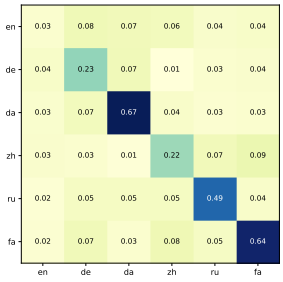
(d) Language

Figure 4: Impact of ablating four neuron subsets on our MUREL test set in Llama-2-7b. Each cell (i, j) shows perplexity (PPL) change on culture j when ablating neurons of language or culture i.

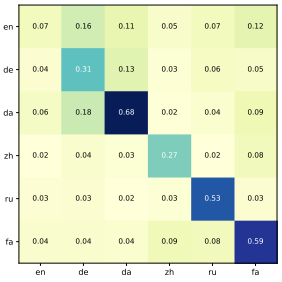


(a) Culture

(b) Pure Culture



(c) Language ∩ Culture



(d) Language

Figure 5: Impact of ablating four neuron subsets on our MUREL test set in Llama-3.1-8b. Each cell (i, j) shows the perplexity (PPL) change on culture j when ablating neurons of language or culture i.

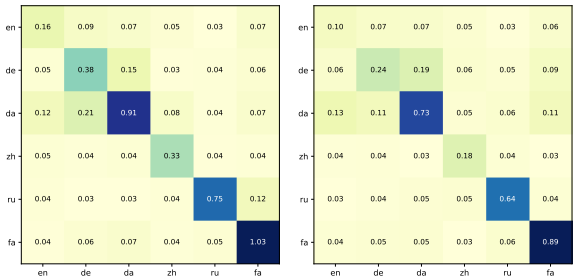
model, we observe a secondary peak in the bottom layers, resulting in a bimodal distribution consistent with previous findings (Tang et al., 2024; Zhao et al., 2024). This may reflect a dual specialization, with early layers capturing lower-level linguistic patterns and top layers encoding higher-level semantics. By contrast, multilingual models (e.g. Llama-3.1-8b) show a more pronounced concentration of both neuron types exclusively in the upper layers, suggesting a more hierarchical organization of semantic information. Notably, pure culture-specific neurons follow a similar pattern but are comparatively sparser across layers.

6.2 Intervention Experiments

To assess the functional roles of identified neuron subpopulations, we conduct ablation experiments by zeroing out: (a) language-specific, (b) culture-specific, (c) pure culture-specific, (d) compound language-and-culture, and (e) randomly selected neurons. We then measure the resulting change in model perplexity on the MUREL dataset.

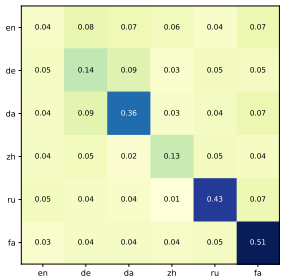
Figure 5 illustrates perplexity changes in the Llama-3.1-8b model after these interventions; diagonal entries reflect effects within the corresponding language or culture, while off-diagonal entries indicate cross-linguistic or cross-cultural impact.

Ablating culture-specific neurons yields the largest increase in perplexity for culturally relevant data, confirming their critical role. Pure culture-specific neurons cause the second-largest perplexity increase, and accounts for about 76.3% of the to-

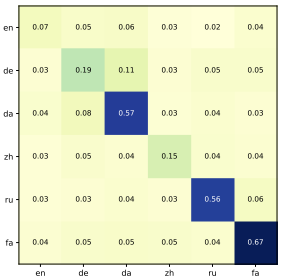


(a) Culture

(b) Pure Culture

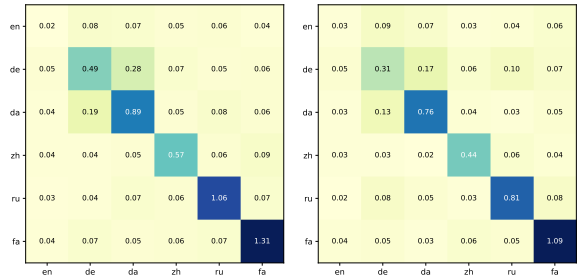


(c) Language ∩ Culture



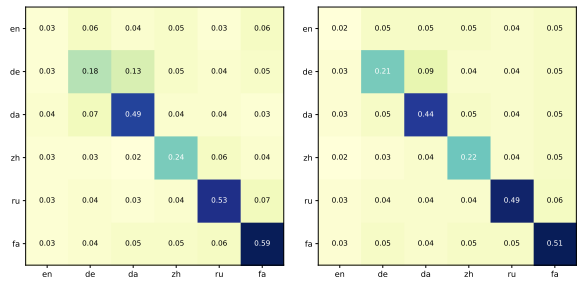
(d) Language

Figure 6: Impact of ablating four neuron subsets on our MUREL test set in Qwen2.5-7b. Each cell (i, j) shows perplexity (PPL) change on culture j when ablating neurons of language or culture i.



(a) Culture

(b) Pure Culture



(c) Language ∩ Culture

(d) Language

Figure 7: Impact of ablating four neuron subsets on our MUREL test set in Gemma-3-12b. Each cell (i, j) shows perplexity (PPL) change on culture j when ablating neurons of language or culture i.

tal effect from ablating all culture-specific neurons. This shows that a large share of cultural knowledge in LLMs is encoded in neurons that are largely independent of language processing.

Crucially, off-diagonal (cross-linguistic and cross-cultural) effects remain consistently minimal, indicating that ablations mainly impact the targeted language or culture. Random neuron ablation has a negligible effect (Figure 8), further emphasizing that the functional roles of identified neuron groups are not due to chance. These patterns hold across all evaluated models, with only minor variation.

7 Discussion and Conclusion

We show that culture-specific neurons—and especially pure culture-specific neurons, which encode cultural knowledge independently of linguistic representations – play a substantial role in shaping model predictions for culturally nuanced content. Although pure culture-specific neurons constitute only about 56.7% of culture-related neurons, their ablation disproportionately increases perplexity, underscoring their functional importance. These results indicate that multilingual language models organize cultural knowledge into specialized neural populations, separable from linguistic encoding. Notably, both language- and culture-specific neurons predominantly reside in upper layers, consistent with hierarchical theories of semantic representation. Thus, our work enhances our understanding of how multilingual models internally represent complex cultural and semantic information.

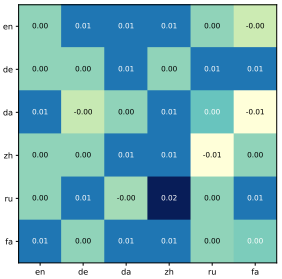


Figure 8: Perplexity changes after randomly ablating neurons in Llama-3.1-8b. Number of ablated neurons per culture matches the average identified per culture.

This work advances our understanding of how cultural information is represented within multilingual language models. Our approach offers a framework for probing the interplay between cultural and linguistic signals in model internals, and facilitates future work on representational structure, identity modeling, or culturally grounded evaluation. Our findings suggest that, like language, culture can be meaningfully localized and examined as a distinct component of model representations. Notably, we did not find any "generic" culture neurons shared across all cultures, i.e., $\bigcap_m \mathbb{P}_m = \emptyset$, indicating that cultural representations are highly specific.

Conclusion We have introduced a methodology to identify and isolate culture neurons in multilingual language models. To facilitate our analyses, we have compiled MUREL, a large and culturally diverse resource. Our results show that cultural knowledge concentrates in specialized neuron populations predominantly localized in upper layers of multilingual language models and that pure culture neurons play a substantial functional role. Ablation experiments demonstrate that each culture is encoded in distinct neural populations with minimal cross-cultural interference.

We invite future research on (i) broadening coverage to additional cultures, languages, and models, (ii) extending localization to attention heads where the present study focuses on the feedforward modules, and (iii) testing sufficiency via activation scaling and steering, as well as evaluating intervention effects on downstream tasks.

8 Limitations

While our study provides new insights into the neural localization of culture in multilingual language models, several limitations remain. First, our analysis is restricted to a small set of open-source models and may not generalize to larger or proprietary LLMs with different architectures or training data. Second, our methodology for identifying culture neurons relies on entropy-based metrics and dataset sampling choices, which may limit our ability to detect more distributed or context-dependent representations. Third, although the MUREL dataset is diverse, it covers only six cultures, potentially omitting important cultural phenomena found in other regions or language families. Finally, our evaluation focuses on neuron ablation and perplexity; future work should include more comprehensive behavioral and downstream assessments to better understand the practical impact of these neurons.

9 Potential Risks

This research analyzes the internal representations of multilingual language models and introduces MUREL, a culturally diverse evaluation dataset. All data used are derived from publicly available and properly credited resources. No private, sensitive, or personally identifiable information was included. While identifying culture-specific neurons may help increase transparency and cultural awareness in language models, it also raises the risk of model manipulation or the reinforcement of cultural stereotypes if misused. Our methodology is intended to advance understanding and fairness in multilingual NLP, not to entrench or amplify cultural biases.

Acknowledgements

This work was supported in part by the Danish Foundation Models project.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2189–2203. ELRA and ICCL.

Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 672–680. Association for Computational Linguistics.

Tosin Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaido, Foteini Liwicki, and Marcus Liwicki. 2022. Potential idiomatic expression (PIE)-English: Corpus for classes of idioms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 689–696. European Language Resources Association.

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a Russian idiom-annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267. Association for Computational Linguistics.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776. Association for Computational Linguistics.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67. Association for Computational Linguistics.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021. Association for Computational Linguistics.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265. Association for Computational Linguistics.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. 2024. Finding safety neurons in large language models. *arXiv preprint arXiv:2406.14144*.

Alina Maria Ciobanu, Shervin Malmasi, and Liviu P. Dinu. 2018. German dialect identification using classifier ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 288–294. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502. Association for Computational Linguistics.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718. Association for Computational Linguistics.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857. Association for Computational Linguistics.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh, and Reza Fahmi. 2020. Irony detection in Persian language: A transfer learning approach using emoji prediction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2839–2845. European Language Resources Association.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023. Comparing styles across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791. Association for Computational Linguistics.

G. Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Cross Cultural Research and Methodology. SAGE Publications.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4919. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd*

Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 752–762. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690. Association for Computational Linguistics.

Shulin Huang, Linyi Yang, and Yue Zhang. 2025. Mceval: A dynamic framework for fair multilingual cultural evaluation of llms. *arXiv preprint arXiv:2507.09701*.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6816, Miami, Florida, USA. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284. Association for Computational Linguistics.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698. Association for Computational Linguistics.

Jeongyeon Kim, Sangho Suh, Lydia Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Designing Interactive Systems Conference*, pages 41–57.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. 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 6919–6971. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

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

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024a. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024b. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, Sky CH-Wang, and Smaranda Muresan. 2023b. NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016–2039. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Shervin Malmasi and Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 164–169. Association for Computational Linguistics.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227. European Language Resources Association (ELRA).

Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221. Association for Computational Linguistics.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Max Müller-Eberstein, Mike Zhang, Elisa Bassignana, Peter Brunsgaard Trolle, and Rob Van Der Goot. 2025. DaKultur: Evaluating the cultural awareness of language models for Danish with native speakers. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 50–58. Association for Computational Linguistics.

Abdullah Mushtaq, Imran Taj, Rafay Naeem, Ibrahim Ghaznavi, and Junaid Qadir. 2025. Worldview-bench: A benchmark for evaluating global cultural perspectives in large language models. *arXiv preprint arXiv:2505.09595*.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790. Association for Computational Linguistics.

Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467. Association for Computational Linguistics.

Claudio Paonessa, Yanick Schraner, Jan Deriu, Manuela Hürlimann, Manfred Vogel, and Mark Cieliebak. 2023. Dialect transfer for Swiss German speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15240–15254. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265. Association for Computational Linguistics.

Prisca Piccirilli, Alexander Fraser, and Sabine Schulte im Walde. 2024. VOLIMET: A parallel corpus of literal and metaphorical verb-object pairs for English–German and English–French. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 222–237. Association for Computational Linguistics.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1132–1144.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772. Association for Computational Linguistics.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. Large language models for Persian-English idiom translation. 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 7974–7985. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: culturally-diverse multilingual visual question answering benchmark. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 11479–11505.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 87–94. Springer.

David Schmidt, Albin Zehe, Janne Lorenzen, Lisa Sergel, Sebastian Düker, Markus Krug, and Frank Puppe. 2021. The FairyNet corpus - character networks for German fairy tales. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 49–56. Association for Computational Linguistics.

Felix Schneider, Sven Sickert, Phillip Brandes, Sophie Marshall, and Joachim Denzler. 2022. Metaphor detection for low resource languages: From zero-shot to few-shot learning in Middle High German. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 75–80. European Language Resources Association.

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. Association for Computational Linguistics.

Nathalie Hau Sørensen and Sanni Nimb. 2025. The Danish idiom dataset: A collection of 1000 Danish idioms and fixed expressions. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 55–63. The University of Tartu Library.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206. Association for Computational Linguistics.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. *Advances in Neural Information Processing Systems*, 37:125019–125049.

Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2023. MoralDial: A framework to train and evaluate moral dialogue systems via moral discussions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2213–2230. Association for Computational Linguistics.

Zhewei Sun and Yang Xu. 2022. Tracing semantic variation in slang. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1313. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane

Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, et al. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, et al. 2025. Qwen2.5 technical report.

Zhichao Yang, Pengshan Cai, Yansong Feng, Fei Li, Weijiang Feng, Elena Suet-Ying Chiu, and Hong Yu. 2019. Generating classical Chinese poems from vernacular Chinese. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6155–6164. Association for Computational Linguistics.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055. Association for Computational Linguistics.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Liutao Liutao, and Deyi Xiong. 2024. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11817–11837. Association for Computational Linguistics.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. 2024. Measuring social norms of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 650–699. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj Sharma, Shilin Qu, Linhao Luo, Ingrid Zukerman, Lay-Ki Soon, Zhaleh Semnani Azad, and Reza Haf. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2712–2722. ACM.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 15296–15319.

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. Association for Computational Linguistics.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776. Association for Computational Linguistics.

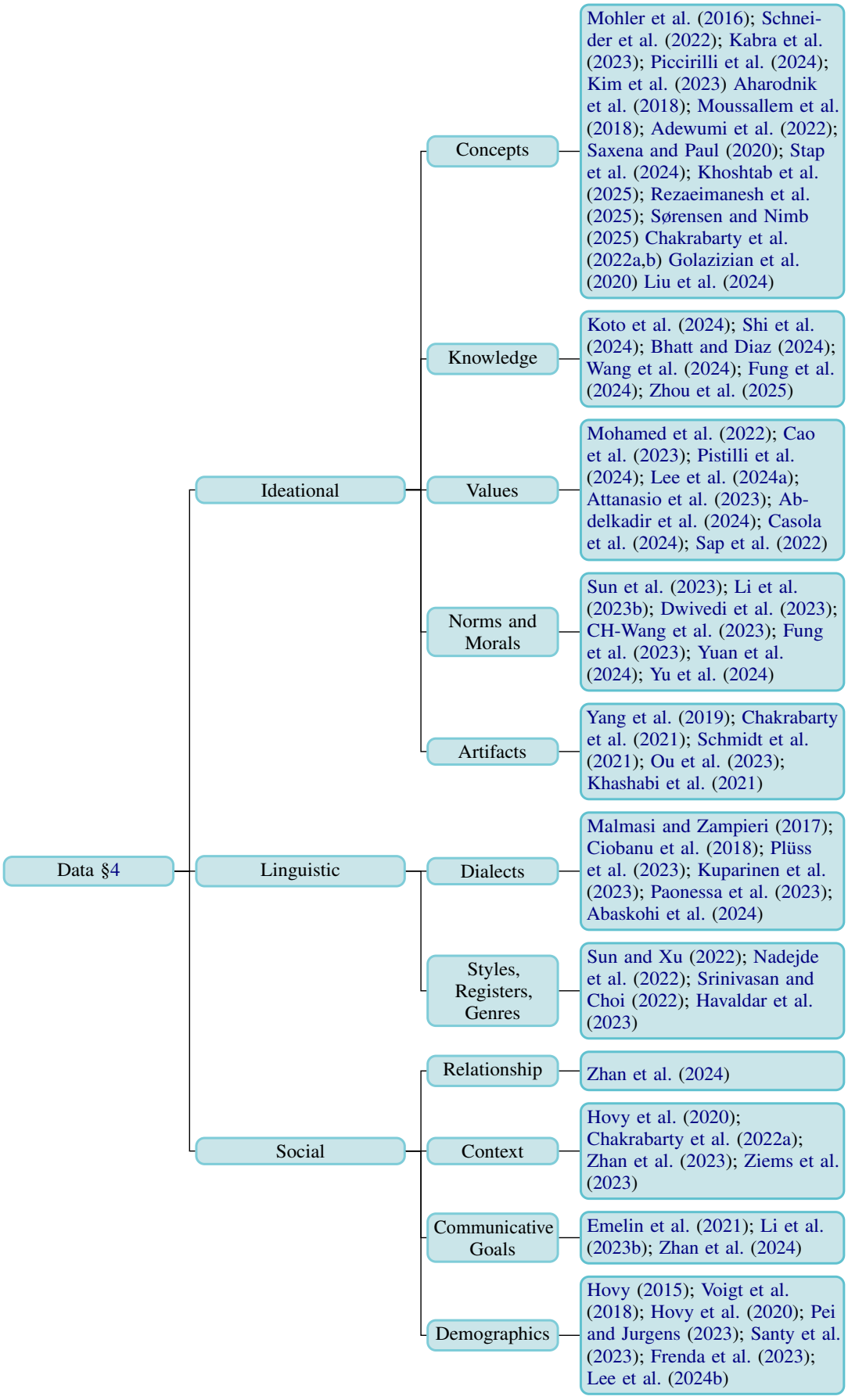


Figure 9: Categorization of cultural data resources in MUREL, with representative references for each category.

A Datasets

We compiled the MUREL dataset, consisting of 69 culturally diverse corpora spanning 6 cultures, with a total of 85.2 million Gemma-3 tokens. All datasets used in this study are publicly available and were used in accordance with their respective open licenses for research purposes only. Table 1 reports the total number of tokens per culture. Figure 9 shows the systematic organization and provides references to all source datasets.

EN	DE	DA	ZH	RU	FA
14,262	19,891	11,383	16,511	12,405	10,769

Table 1: Total number of tokens per culture in MUREL (in thousands).

B Models

For our investigation, we select four transformer-based language models.

B.1 Llama 2

Llama 2[5] is a 7-billion-parameter decoder-only transformer model developed by Meta. It consists of 32 layers and 352,256 neurons. It was trained on a corpus comprising approximately 2 trillion tokens of publicly available online data. While Llama 2 supports text generation in English and 27 other languages, its training data is predominantly English, which may affect performance in less-represented languages.

B.2 Llama 3.1

Llama 3.1[6] is an 8-billion-parameter multilingual model from Meta, with 32 layers and 458,752 neurons. It was trained on diverse text corpora. The model consists of stacked transformer layers, each comprising self-attention and feedforward MLP components. Llama 3.1 is optimized for computational efficiency and supports a wide range of languages, making it a strong candidate for evaluating multilingual transfer performance.

B.3 Gemma 3

Gemma 3[7] is a 12-billion-parameter transformer-based model developed by Google, with 48 layers and 737,280 neurons. It is part of the Gemma family of lightweight, open models built from the same research and technology used to create Gemini. Gemma 3 models are multimodal and have a large, 128K context window, multilingual support in over 140 languages, and are available in more sizes than previous versions.

B.4 Qwen 2.5

Qwen 2.5[8] is a 7-billion-parameter decoder-only transformer model developed by Alibaba Cloud, comprising 28 layers and 100,352 neurons. The Qwen2.5-7B model was trained on a substantial corpus of 18 trillion tokens, significantly expanding upon the 7 trillion tokens used in its predecessor, Qwen2. The model supports over 29 languages, making it a robust choice for multilingual applications.

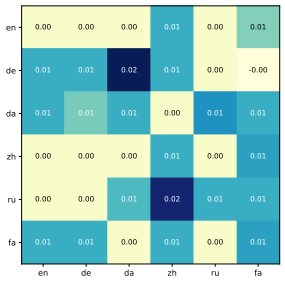


Figure 10: Perplexity changes after randomly ablating neurons in Llama-2-7b. Number of ablated neurons per culture matches the average identified per culture.

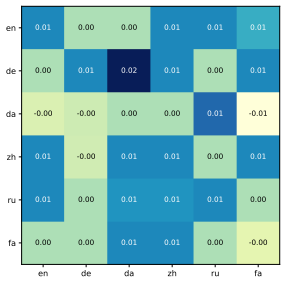


Figure 11: Perplexity changes after randomly ablating neurons in Qwen2.5-7b. Number of ablated neurons per culture matches the average identified per culture.

C Additional Random Ablations

Figures 10, 11, and 12 show the random ablations for Llama-2-7b, Qwen2.5-7b, and Gemma-3-12b, respectively.

[5] https://huggingface.co/meta-llama/Llama-2-7b

[6] https://huggingface.co/meta-llama/Llama-3.1-8B

[7] https://huggingface.co/google/gemma-3-12b-pt

[8] https://huggingface.co/Qwen/Qwen2.5-7B

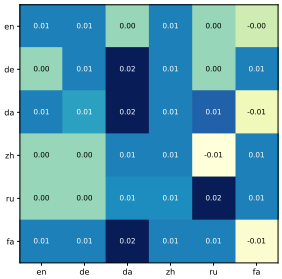
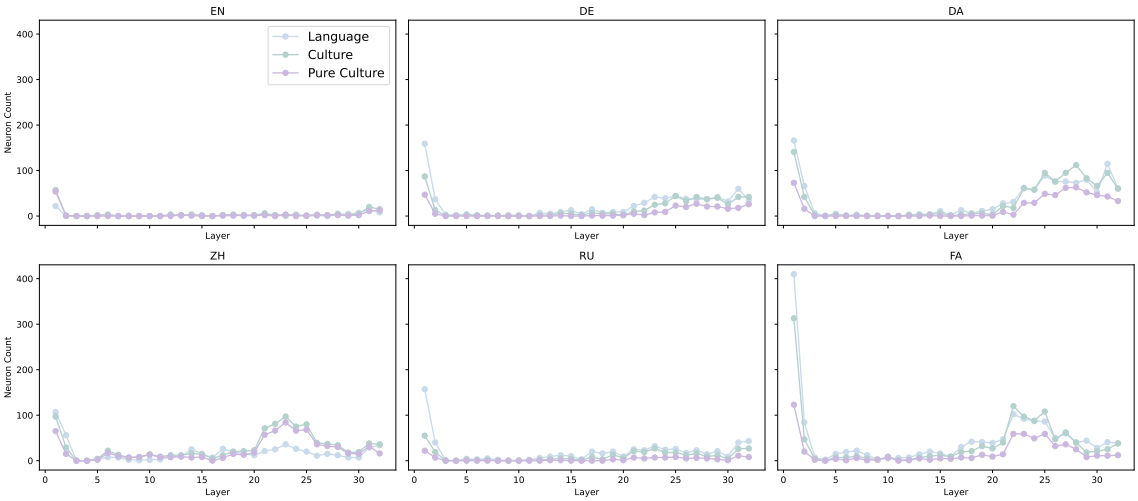


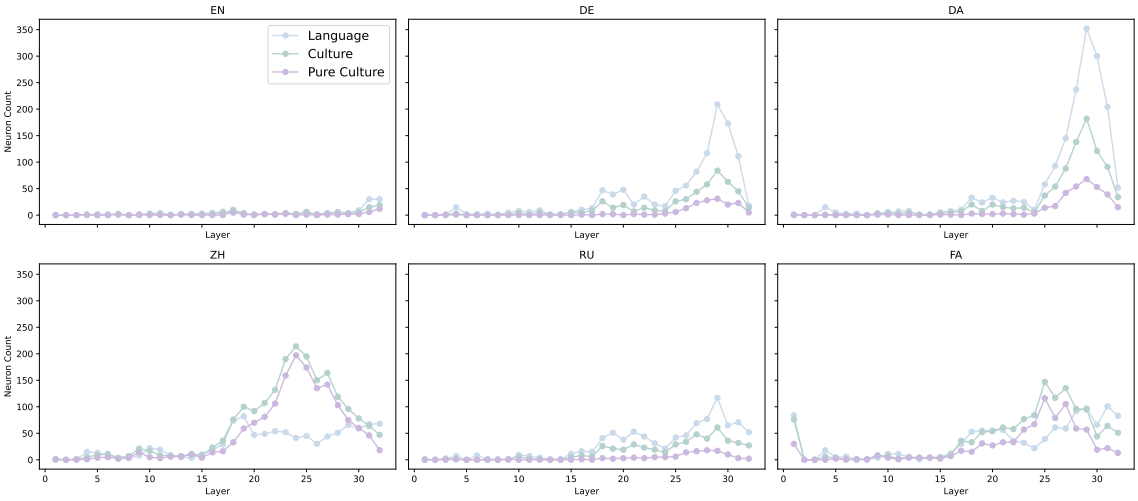
Figure 12: Perplexity changes after randomly ablating neurons in Gemma-3-12b. Number of ablated neurons per culture matches the average identified per culture.

D Computational Infrastructure

All experiments, including neuron activation analysis and ablation interventions, were conducted using pretrained models without any additional training or fine-tuning. Computations were performed on a single NVIDIA V100 GPU per experiment. Across all models, the total computational budget did not exceed 280 GPU hours.



(a) Llama-2-7b



(b) Llama-3.1-8b

Figure 13: Layer-wise distribution of language, culture, and pure culture neurons for each language, visualized for (a) Llama-2-7b and (b) Llama-3.1-8b.