BEYOND TRAINING FOR CULTURAL AWARENESS: THE ROLE OF DATASET LINGUISTIC STRUCTURE IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The global deployment of large language models (LLMs) has raised concerns about cultural misalignment, yet the linguistic properties of fine-tuning datasets used for cultural adaptation remain poorly understood. We adopt a dataset-centric view of cultural alignment and ask which linguistic properties of fine-tuning data are associated with cultural performance, whether these properties are predictive prior to training, and how these effects vary across models. We compute lightweight linguistic, semantic, and structural metrics for Arabic, Chinese, and Japanese datasets and apply principal component analysis separately within each language. This design ensures that the resulting components capture variation among datasets written in the same language rather than differences between languages. The resulting components correspond to broadly interpretable axes related to semantic coherence, surface-level lexical and syntactic diversity, and lexical or structural richness, though their composition varies across languages. We fine-tune three major LLM families (LLaMA, Mistral, DeepSeek) and evaluate them on benchmarks of cultural knowledge, values, and norms. While PCA components correlate with downstream performance, these associations are strongly model-dependent. Through controlled subset interventions, we show that lexical-oriented components (PC3) are the most robust, yielding more consistent performance across models and benchmarks, whereas emphasizing semantic or diversity extremes (PC1–PC2) is often neutral or harmful. (Reproducibility notebook: `https://anonymous.4open.science/r/dataset_quality_demo-C7C9/demo_subset_selection.ipynb`.)

1 INTRODUCTION

LLMs have shown remarkable progress across diverse natural language processing (NLP) tasks. However, their performance often falls short in cross-cultural settings, where linguistic variation, cultural norms, and local knowledge shape how users interpret and engage with model outputs. Cultural alignment, which ensures that LLMs understand and reflect the values, norms, and nuances of the user groups interacting with it Masoud et al. (2023), is therefore essential for building inclusive, globally applicable AI systems. Recent work suggests that dataset quality profoundly influences model performance Zhou et al. (2023); Alshahrani et al. (2023), yet cultural datasets themselves remain understudied from a linguistic and structural perspective. Dominant approaches to cultural alignment predominantly focus on value congruence, relying heavily on survey-derived benchmarks (e.g., World Values Survey) or synthetic QA pairs with limited attention to the linguistic structure of the underlying training data. However, these methods often overlook the *linguistic* dimension, neglecting how cultural nuances are encoded in the structural, semantic, and stylistic properties of the training data itself, and whether such properties can be assessed independently of model training.

Despite this progress, cultural datasets remain under-represented for many non-Western languages Pawar et al. (2025). Little is known about how their linguistic properties influence cultural alignment, and prior work suggests that different model architectures may respond differently to the same training data Yauney et al. (2023); Zhang et al. (2025). We therefore ask whether cultural fine-tuning datasets exhibit consistent, measurable patterns in semantic content, lexical diversity, and stylistic variability that can be quantified prior to training, and whether such properties are predictive of, and actionable for, downstream cultural alignment across different model families. To answer these questions, we (i) characterize multilingual datasets using lightweight linguistic metrics and PCA, (ii) test associations between PCA dimensions and cultural benchmark performance across models, and (iii) probe actionability through controlled high/low/random subset finetuning.

Our contributions are threefold:

1. We present a *dataset-centric methodology* that quantifies linguistic, semantic, and structural properties of cultural datasets, reduces them via PCA, and links these components to downstream cultural alignment performance.
2. To our knowledge, we conduct the *first* cross-lingual empirical study examining Arabic, Chinese, and Japanese datasets across three LLM families, revealing that correlations between dataset properties and cultural performance vary substantially by language and model architecture.
3. We evaluate the *predictive* utility of PCA-derived linguistic dimensions for dataset assessment, including controlled subset-based interventions that test whether these signals remain informative under fixed training conditions.

Overall, our results indicate that pre-training linguistic properties of datasets can be informative for cultural alignment, but their effects are strongly model-dependent rather than universal. This motivates model-aware, dataset-centric strategies for multilingual cultural awareness in LLMs.

2 Related Work

Prior work on cultural alignment has introduced benchmarks and datasets, often survey-based or value-oriented, that compare model outputs to human responses (e.g., AlKhamissi et al. 2024; Masoud et al. 2023). While effective for evaluation, these approaches typically do not analyze the linguistic or structural properties of the datasets themselves, nor how such properties relate to downstream cultural alignment. Complementary work has examined cultural bias and representational gaps in pretraining and post-training corpora (Naous et al., 2023; Alkhowaiter et al., 2025), as well as methods for curating culturally diverse datasets (Li et al., 2024). However, these studies do not identify which dataset-level linguistic properties are predictive of cultural alignment performance. In contrast, our work adopts a dataset-centric perspective, quantifying linguistic properties of multilingual cultural datasets and linking them to alignment outcomes across models, enabling assessment of dataset utility prior to fine-tuning. Additional adjacent related work is discussed in Appendix B.

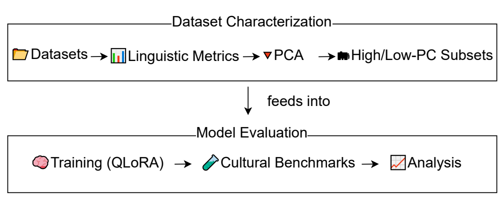


Figure 1: Methodology

3 Methodology

This work presents a dataset-centric methodology for analyzing how linguistic and structural properties of fine-tuning datasets relate to downstream cultural awareness in LLMs. Our approach is model-agnostic and operates entirely at the dataset level, allowing dataset characteristics to be quantified *ex-ante* (prior to training), rather than relying on *post-hoc* analysis of model behaviour. The pipeline consists of two stages (Figure 1): (1) dataset characterization (Steps 1–3), where linguistic metrics are computed, reduced via PCA, and used to define controlled dataset subsets; and (2) model evaluation (Steps 4–6), where LLMs are fine-tuned and assessed on cultural alignment benchmarks. The steps are detailed below. Each step is designed to test whether dataset-level linguistic structure captures systematic differences between datasets, correlates with downstream cultural performance, and supports systematic data selection that is more informative than random subset baselines.

Step 1: Linguistic Feature Extraction For each dataset, we compute a set of lightweight linguistic, semantic, and structural metrics designed to capture complementary aspects of language use that may plausibly affect cultural reasoning, including lexical richness, surface-level diversity, semantic similarity, and structural cohesion (see Table 1).

Table 1: Summary of text evaluation metrics. Diversity metrics include Distinct-1/2 (Li et al., 2016) and Self-BLEU (Papineni et al., 2002). Lexical richness is measured using TTR (Johnson, 1944), MATTR (Covington & McFall, 2010), and HDD Wu (1993); McCarthy & Jarvis (2010), MTLD (McCarthy & Jarvis, 2010)). Semantic similarity is computed using cosine similarity and TF-IDF representations (Salton et al., 1975; Salton & Buckley, 1988; Reimers & Gurevych, 2019). Clustering structure is assessed using K-means (McQueen, 1967) and the silhouette score (Rousseeuw, 1987).

Category	Metrics	Key Focus
Diversity	Distinct-1/2, Self-BLEU	Measures n-gram uniqueness and cross-sample repetition.
Lexical Richness	TTR, MATTR, HDD, MTLD	Captures vocabulary breadth and distribution stability.
Semantic Similarity	Cosine Similarity, TF-IDF	Evaluates meaning proximity using vector representations.
Clustering Structure	Silhouette Score, K-means	Reflects semantic cohesion and separation in embedding space.

All metrics are computed at the dataset level, treating each dataset as a collection of samples whose aggregated properties reflect its linguistic profile. To ensure comparability across datasets of different scales, we randomly sample 1,000 examples per dataset. Additional pilot experiments with larger samples and repeated random draws showed minimal variation, supporting the stability of this sampling strategy. Because languages exhibit different morphological and statistical properties, we normalize features separately for each language by rescaling each metric to zero mean and unit variance. This ensures that PCA compares datasets based on their relative linguistic properties within a language, rather than grouping datasets by language due to differences in metric magnitude (e.g., stemming from morphology or tokenization).

Step 2: PCA-Based Dimensionality Reduction To further compress the 10 specified linguistic metrics into a small number of interpretable dataset-level descriptors, we apply PCA separately for each language. PCA combines correlated linguistic metrics into a few continuous components, each assigning a score to every dataset that reflects how strongly it exhibits a particular combination of linguistic properties. We represent each dataset by its scores along the first three principal components (PC1–PC3), which capture most of the variance in the linguistic metrics. The resulting PCA scores serves as the basis for subsequent analyses that examine associations with downstream cultural performance and test whether these dimensions can guide dataset selection.

Step 3: Dataset Profiling and Subset Construction While PCA reveals descriptive structure, it does not establish whether these dimensions are useful for guiding training decisions. To test the practical relevance of PCA-derived signals, we construct controlled dataset subsets that approximate movement along each principal component at the sample level. Specifically, for each dataset and each principal component, we identify the linguistic metric with the highest absolute loading on that component and use it as a proxy to rank individual samples. We then construct equal-sized fine-tuning subsets by selecting samples with high or low values of this proxy metric. We additionally construct a random subset of equal size as a baseline. All subsets are matched in number of examples to isolate linguistic effects from dataset scale. By manipulating dataset composition in this way, we test whether dataset-level associations persist under controlled subset selection or disappear when data is re-sampled.

Step 4: Model-Agnostic Fine-Tuning Setup All fine-tuning experiments start from the same base LLM checkpoint. We independently fine-tune this base model on each full dataset and on each PCA-based or random subset constructed in Step 3, using an identical training configuration throughout. No model is fine-tuned sequentially on multiple datasets or subsets. This produces a collection of models that differ only in the data used for fine-tuning, allowing us to isolate the effect of dataset composition while controlling for model architecture and optimization settings.

Step 5: Cultural Evaluation Each fine-tuned model is evaluated on a set of cultural benchmarks covering three categories: cultural knowledge, cultural values, and cultural norms. These benchmarks provide model-level performance scores across distinct aspects of cultural alignment, enabling systematic comparison across datasets, subsets, and model families.

Step 6: Linking Dataset Properties to Model Performance Finally, we analyze whether dataset-level linguistic structure is associated with downstream cultural behavior by computing correlations between (i) each dataset's PCA coordinates (PC1–PC3) and (ii) the performance of the corresponding fine-tuned model on cultural benchmarks. These correlations are used to assess predictive association, not causality, and motivate the subset-based intervention experiments that follow.

4 Experiments

4.1 Setup

Languages and Datasets We conduct experiments across Arabic, Chinese, and Japanese, using between 9 and 13 post-training datasets per language. To ensure broad coverage of linguistic and cultural variation, the selected datasets span diverse sources and domains such as exams, social media, news, instruction-tuning corpora, general web collections, and human-curated or annotated resources. This diversity enables us to capture a wide range of linguistic structures, cultural registers, and communicative styles. These include:

Arabic: MultiNativQA Hasan et al. (2024), ArabicMMLU Koto et al. (2024), Aya Singh et al. (2024), CIDAR Alyafeai et al. (2024), Open-ArabicaQA Abdallah et al. (2024), The Ultimate Arabic News Al-Dulaimi (2022), DAWQAS Ismail & Homsi (2018), Al Jazeera News Articles ArbML (2023a), and Arsen-20 ArbML (2023b)

Japanese: JaQuAD So et al. (2022), Japanese WordNet 2.0 Bond & Kuribayashi (2023), JcommonsenseMorality Takeshita & Araki (2023), Ichikara Instruction All msfm (2023), Dolly 15k-JA Kunishou (2023), AIO Instruction Dataset Project (2023), Core Vocabulary Simplification Corpus Katsuta & Yamamoto (2018), Wikimedia Dumps Wikimedia Foundation (2024a), Japanese Wikinews Wikimedia Foundation (2024b), and Aozora Bunko (AozoraTxt) levelevel (2023).

Chinese: COIG-PC, Chinese_Traditional, Douban, Xiaohongshu (XHS), Human_Value, RuozhiBa, Exam, SegmentFault, Zhihu, LogiQA, Wikihow, Wiki, Finance Bai et al. (2024)

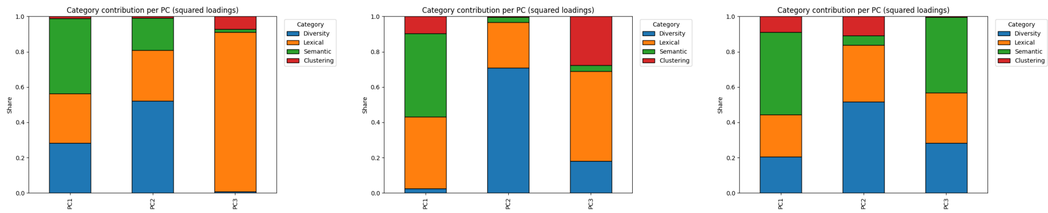
Models Llama-3.2-3B-Instruct Grattafiori et al. (2024), Mistral-7B-Instruct-v0.3 Jiang et al. (2023), and DeepSeek-R1-Distill-Qwen-7B Guo et al. (2025).

Evaluation Framework Models are evaluated across three cultural alignment categories: Cultural Knowledge, Cultural Values, and Cultural Norms. Benchmarks include VSM13, World Values Survey (WVS), CulturalBench, EXAMs, ACVA, CIDAR-EVAL, and additional culture-related evaluation datasets relevant to each language (full list in Appendix D).

Processing and Fine-Tuning Protocol For linguistic metric computation, each dataset was randomly sampled to 1,000 examples. For model training, datasets with ¡30k examples were used fully, larger datasets were capped at 30k samples, split into 80% train / 10% validation / 10% test. All additional fine-tuning configurations, including sequence length, optimizer settings, learning rate schedule, batch size, training steps, and hardware setup, are documented in Appendix C.

4.2 Linguistic Feature Analysis

As described in Section 3, we compute linguistic, semantic, and structural metrics for all datasets and apply PCA separately for each language to capture within-language variation. Here, we analyze the resulting PCA components to compare how Arabic, Japanese, and Chinese datasets differ along the major axes of linguistic variation.



(a) Arabic (b) Japanese (c) Chinese

Figure 2: Category-level contributions of diversity, lexical, semantic, and clustering metrics to principal components PC1–PC3 for Arabic, Japanese, and Chinese.

4.2.1 Principal Components Capture Meaningful Linguistic Structure

For each language, we apply PCA to the matrix of dataset-level linguistic metrics, where each dataset corresponds to one observation and each metric to one feature. The first three principal components (PCs) capture coherent groupings of linguistic features and explain most of the dataset-level variance. PC1 explains 41–53% of variance across languages, with PC2 and PC3 contributing a further 18–33% and 9–16%, respectively; additional components explain comparatively little variance. Category-level contribution plots (Fig. 2) show that, within each language, individual principal components are dominated by subsets of linguistic metric categories rather than uniform mixtures. The dominant categories differ across components and languages, indicating that PCA captures language-specific combinations of linguistic properties rather than a single shared linguistic dimension. This motivates examining whether different components play distinct roles in downstream cultural alignment.

PC1: Semantic-Dominant Dimension PC1 captures semantic coherence and similarity, often combined with surface-level diversity (Arabic) or lexical richness (Japanese, Chinese). This component explains the largest share of variance in every language (approximately 41–53%), reflecting its aggregation of multiple high-level linguistic signals associated with meaning-dense and conceptually structured datasets.

PC2: Diversity + Lexical Dimension PC2 reflects surface variability and topic spread, capturing differences in lexical diversity, distributional heterogeneity, and stylistic variation. This component cuts across dataset sources and emphasizes differences in how broadly language use is distributed rather than domain-specific content.

PC3: Secondary Lexical/Semantic Structure PC3 captures more localized, language-specific structure. In Arabic, it is driven primarily by lexical richness metrics (e.g., MTLD, HDD); in Japanese, it combines lexical and clustering signals that separate more homogeneous from heterogeneous corpora; and in Chinese, it reflects secondary semantic structure dominated by similarity metrics. Compared to PC1 and PC2, PC3 aggregates a narrower set of linguistic signals, emphasizing lexical and stylistic variation rather than broad semantic or topical structure.

4.2.2 Dataset Interpretation Through PCA Dimensions

By projecting datasets into PCA space, datasets exhibit separations that reflect differences in their linguistic composition. Below, we describe how major dataset families are distributed along each component, as illustrated in Fig. 3.

PC1: Datasets with Higher Semantic Density PC1 separates datasets according to how much explicit semantic reasoning they contain. Datasets built around knowledge-intensive QA datasets (e.g.,such as NativQA, MBZUAI Arabic MMLU, OpenArabicQA, JCommonsense, COIG-PC, and LogiQA), consistently achieve high PC1 scores. These sources contain meaning-dense prompts and well-structured QA pairs, which increases semantic similarity metrics and drives strong PC1 loadings. By contrast, large news and encyclopedic corpora, including *Ultimate Arabic News*, *Aljazeera*, *Wikipedia-30k*, *Wikinews-5k*, *Wikihow*, and *Wiki*, tend to score low on PC1. Their formulaic reporting style, repetitive phrasing, and narrower topical variation produce lower semantic variability, placing them at the lower end of the PC1 axis.

PC2: Dataset Variation Along Surface-Level Properties PC2 does not align cleanly with any dataset source. Datasets from news, QA, instructional, and social-media origins appear throughout this dimension with no consistent pattern. This suggests that PC2 captures surface-level variation, such as differences in token distribution, topical breadth, or stylistic alternation, that cuts across genres rather than characterizing any particular dataset family.

PC3: Datasets with Greater Lexical and Stylistic Variation High PC3 scores are associated with datasets that contain human-authored, open-ended, and stylistically diverse text. These include *Aya* (instructional, human-written), *CIDAR* (free-text survey responses), *JCommonsense* (crowd-sourced QA), *Ichikara* (human-curated instructions), *Douban* (social-media dialogue), and *COIG-PC* (conversation-style preferences). Some news and encyclopedic datasets, such as *Ultimate Arabic News*, *Wikinews*, and *Wiki*, also appear high on PC3. Although stylistically formulaic, these sources cover a wide range of topics and entities, resulting in a broad vocabulary and higher lexical richness, which contributes to their elevated PC3 scores. Across these examples, datasets that provide varied vocabulary or unconstrained language use exhibit strong PC3 loadings, regardless of whether their style is structured or conversational.

Overall, the PCA projections provide a compact, language-specific view of how datasets differ in their linguistic composition. Rather than revealing shared dimensions across languages, the PCA space offers a descriptive organization of datasets within each language, which we use in subsequent analyses to examine relationships with cultural alignment performance.

4.3 Dataset Linguistic Structure and Its Impact on Cultural Alignment

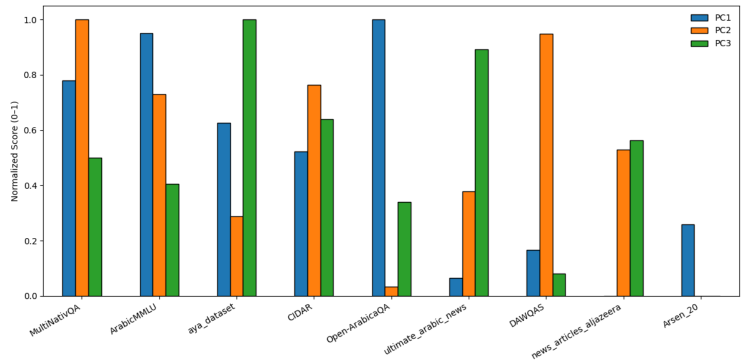
This experiment examines whether dataset-level linguistic structure, as captured by PCA, is systematically associated with downstream cultural-alignment performance across languages and model families. For each dataset analyzed in Section 4.2, we fine-tune the same base LLM using a fixed training configuration, yielding one model per dataset. Each model is then evaluated on cultural benchmarks grouped into three categories: *Cultural Knowledge*, *Cultural Values*, and *Cultural Norms*. We compute Pearson correlations between (i) each dataset's PCA coordinates (PC1–PC3) and (ii) the performance of the corresponding fine-tuned model on cultural alignment benchmarks, using the task-specific evaluation metrics defined in Appendix D. Importantly, these correlations are used to assess associative relationships, not to establish causality or actionability. Whether PCA-derived dimensions can reliably guide dataset selection is tested separately via controlled subset interventions in Section 4.4. Heatmaps for Arabic, Japanese, and Chinese are shown in Figs. 5, 6, and 7.

4.3.1 PCA Components Exhibit Systematic but Model-Dependent Associations

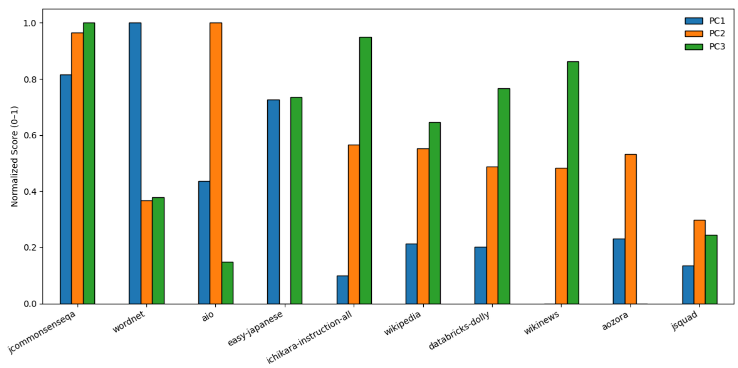
Across all three languages, most benchmark categories exhibit moderate or strong correlations with at least one PCA component, indicating that dataset-internal linguistic structure is systematically associated with cultural-alignment outcomes. However, the identity of the relevant component varies substantially across models, benchmarks, and languages, suggesting that no single PCA dimension acts as a universal predictor.

Arabic: Arabic benchmarks exhibit varied associations with PCA components. Cultural knowledge and values tasks are often most strongly associated with PC1 or PC2, but the identity of the dominant component depends on the model architecture. For example, CultureAtlas is most strongly associated with PC1 for LLaMA (0.85), whereas EXAMs is most strongly associated with PC2 for Mistral (0.82). Cultural values benchmarks show similarly strong but model-dependent associations (e.g., WorldValuesBench with PC1 for Mistral (-0.77) and ACVA with PC2 for DeepSeek (0.78)). Normative benchmarks such as CIDAR-Eval also exhibit shifting associations across models, with PC2 most strongly associated with performance for LLaMA (0.71) and PC3 for DeepSeek (0.60). While PC3 appears among positive associations for some model–task pairs, it does not consistently dominate across benchmarks, indicating that no single PCA component serves as a universal predictor.

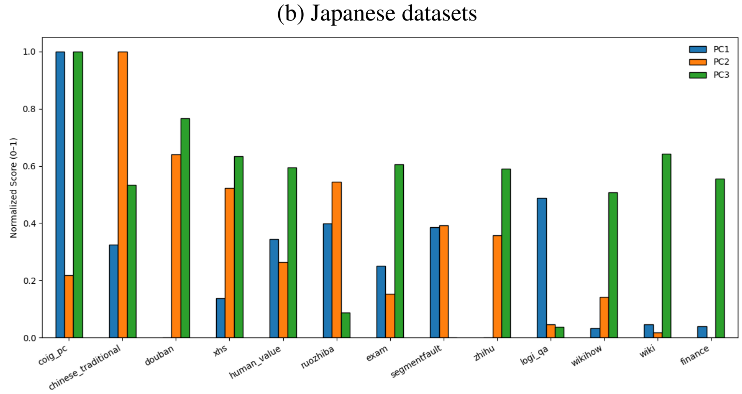
Japanese: In Japanese, cultural knowledge benchmarks tend to exhibit their strongest associations with PC1 (e.g., CulturalBench-Easy: LLaMA -0.66, Mistral -0.68, DeepSeek -0.43), suggesting sensitivity to semantic coherence. In contrast, cultural values and norms benchmarks align with



(a) Arabic datasets



(b) Japanese datasets



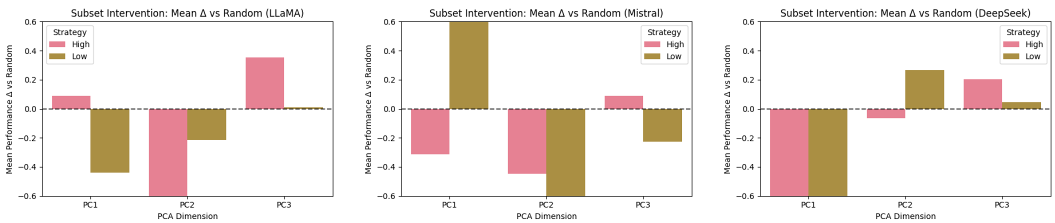
(c) Chinese datasets

Figure 3: Normalized PCA scores for all datasets in each language, projected onto the three main components (semantic relevance, diversity, lexical richness).

different components depending on the model, with observed shifts between PC1, PC2, and PC3 across architectures. These patterns suggest that Japanese cultural performance is influenced by multiple linguistic dimensions, whose relevance is modulated by model design.

Chinese: Chinese benchmarks display similarly distributed associations across PCA components. Cultural knowledge and values tasks align with different PCs depending on the model, while normative benchmarks exhibit moderate but nonzero correlations across components. For example, WorldValuesBench is most strongly associated with PC1 for DeepSeek, while the same benchmark is associated with PC2 for Mistral. As in Arabic and Japanese, PC1 and PC2 show inconsistent behavior across architectures, while PC3 appears more selectively associated with performance for certain models and tasks.

Notably, the same benchmark often aligns with different PCA components depending on model architecture (e.g., VSM13 aligns with PC2 for LLaMA but with PC1 for Mistral, while WorldValuesBench aligns with PC1 for DeepSeek but with PC2 for Mistral), indicating that different models exploit distinct linguistic aspects of the same dataset during fine-tuning. Overall, these results show that while dataset-level linguistic structure is informative for cultural alignment, the identity and direction of relevant components vary substantially across languages, benchmarks, and model families. This variability motivates controlled intervention experiments to test whether any of these associations are actionable for dataset selection.



(a) LLaMA (b) Mistral (c) DeepSeek

Figure 4: Mean performance difference (Δ= subset - random) for High-PC and Low-PC subsets of equal size across PC1–PC3 and three model families, averaged over all base datasets and evaluation metrics. Zero indicates parity with random selection; positive values denote improvement and negative values degradation. PC1-PC2 rarely outperform random sampling, while PC3 shows more consistent, but model-dependent, gains.

4.4 Subset Validation

Correlation alone does not establish whether dataset-level linguistic dimensions are actionable for training. We therefore perform a controlled subset intervention, directly manipulating dataset composition while holding the model, optimization procedure, and subset size fixed. We focus on Arabic because it has the largest and most diverse collection of fine-tuning datasets in our study, spanning a wide range of sources and domains. We identify the five datasets with the largest absolute scores along each principal component (PC1–PC3). Within each dataset, we assign sentences a proxy PC score by projecting their linguistic feature vectors onto the corresponding PCA direction. Using this score, we construct three size-matched subsets ($\approx$ 2k examples): a *High-PC* subset (upper tail), a *Low-PC* subset (lower tail), and a *Random* subset. Each subset independently fine-tunes the same base models (*LLaMA*, *Mistral*, *DeepSeek*) under identical hyperparameters, followed by evaluation on cultural benchmarks.

PC1 and PC2 interventions test whether emphasizing or suppressing semantic structure and distributional diversity improves alignment, while PC3 probes lexical and stylistic variation. Random subsets serve as a strong baseline that controls for subset size while implicitly reducing redundancy, allowing gains or failures to be attributed to composition rather than scale. Figure 4 reports the mean performance difference (Δ) relative to Random for each strategy and PC, averaged across all evaluated datasets and metrics; positive values indicate improvement over random sampling.

LLaMA: LLaMA exhibits limited but structured sensitivity to PCA-guided subset selection. Along PC1, High-PC subsets provide small positive gains over Random, while Low-PC subsets consistently underperform. PC2 interventions are uniformly harmful, with both High-PC and Low-PC subsets degrading performance relative to Random. In contrast, PC3 shows the clearest positive signal: High-PC3 subsets yield consistent improvements, while Low-PC3 remains near-neutral. Overall, LLaMA benefits selectively from lexical-level variation (PC3), while semantic and diversity-driven extremes offer limited or negative utility.

Mistral: Mistral displays a strongly directional response to PCA-guided subset selection. Along PC1, *Low-PC* subsets substantially outperform both High-PC and Random, while High-PC1 leads to clear degradation. This indicates that fine-tuning on semantically extreme subsets—particularly those emphasizing high semantic density—induces harmful distributional shifts for Mistral. PC2 interventions are consistently detrimental: both High-PC2 and Low-PC2 subsets produce large performance drops relative to Random, suggesting that aggressive manipulation of diversity-related properties is poorly tolerated. In contrast, PC3 emerges as the most reliable dimension. High-PC3 subsets consistently outperform Random, while Low-PC3 underperform, indicating that increased lexical and stylistic variation provides a stable and beneficial training signal for Mistral.

DeepSeek: DeepSeek follows a distinct pattern. For PC1, both High-PC and Low-PC subsets substantially underperform Random, suggesting that semantic extremes are broadly misaligned with the model. Along PC2, Low-PC subsets outperform High-PC and Random, while High-PC2 leads to degradation. PC3 produces consistent gains: High-PC3 yields the strongest improvements, while Low-PC3 provides smaller but still positive effects. Thus, for DeepSeek, PC1 acts as a negative selection signal, PC2 favors reduced diversity, and PC3 provides a robust positive intervention.

Cross-model Comparison: Across models, PCA-guided subset selection exhibits clear *model-dependent* effects. PC3 (lexical and stylistic variation) provides the most consistent and transferable signal, improving performance for all three models when selected in the appropriate direction. A plausible explanation is that PC3 emphasizes surface-level and stylistic diversity without strongly altering the underlying semantic distribution, allowing models to benefit from increased lexical coverage while remaining close to their pretraining regime. In contrast, PC1 and PC2 are more fragile: subsets emphasizing semantic density or distributional extremes can induce larger shifts in the training distribution, which different architectures tolerate unevenly. As a result, these interventions frequently degrade performance and show limited cross-model agreement. Importantly, High- and Low-PC subsets are not symmetric—the direction of intervention along a PCA axis matters as much as the axis itself.

Overall, these results demonstrate that PCA-derived linguistic dimensions can inform subset construction, but only when interpreted in a model- and direction-aware manner. PC3 offers the most stable intervention signal, while aggressive manipulation along PC1 and PC2 often harms alignment. Rather than indicating that certain linguistic properties are universally beneficial or harmful, the findings highlight that increasing or decreasing the same property can produce qualitatively different outcomes depending on the model.

5 Conclusion

This work examines how *dataset-level linguistic properties* relate to cultural alignment in large language models. Across 160 fine-tuned models spanning three languages and three model families, we combine language-specific PCA, correlation analysis, and controlled subset interventions to characterize linguistic structure in cultural datasets and assess its downstream impact. Knowledge-intensive QA datasets align with semantic coherence (PC1), news and encyclopedic corpora cluster in low-semantic regions, and human-authored or conversational data exhibit higher lexical richness (PC3).

Subset-based interventions show that emphasizing semantic density (PC1) or surface diversity (PC2) often leads to unstable or negative effects, whereas the lexical–stylistic dimension (PC3) provides a more robust and transferable signal. Importantly, all effects are strongly model-dependent, indicating that linguistic signals must be interpreted in an architecture-aware manner. While prior work has shown that data diversity can improve model performance and robustness in supervised fine-tuning settings Chen et al. (2024); Pang et al. (2024); Zhou et al. (2023), our results suggest that for cultural awareness tasks, increased semantic density or surface-level diversity is not uniformly beneficial. Instead, dataset composition and model-specific interactions play a more central role.

Practical Implication: For culturally-aware fine-tuning, we recommend: (i) prioritizing dataset composition, as PCA-guided subset selection yields measurable performance differences under fixed training conditions; (ii) treating semantic and diversity signals (PC1, PC2) with caution due to their instability across architectures; and (iii) favoring high-PC3 subsets, which provide the most consistently positive signal among the examined components. Overall, PCA-derived linguistic dimensions are not universal quality metrics, but practical tools for probing and shaping dataset composition under controlled conditions.

References

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. Arabicaqa: A comprehensive dataset for arabic question answering, 2024.

Ahmed Hashim Al-Dulaimi. *Ultimate Arabic News Dataset*. 05 2022. doi: 10.17632/jz56k5wxz7.1.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*, 2024.

Mohammed Alkhowaiter, Norah Alshahrani, Saied Alshahrani, Reem I Masoud, Alaa Alzahrani, Deema Alnuhait, Emad A Alghamdi, and Khalid Almubarak. Mind the gap: A review of arabic post-training datasets and their limitations. *arXiv preprint arXiv:2507.14688*, 2025.

Saied Alshahrani, Norah Alshahrani, Soumyabrata Dey, and Jeanna Matthews. Performance implications of using unrepresentative corpora in arabic natural language processing. In *Proceedings of ArabicNLP 2023*, pp. 218–231, 2023.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran AQ Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, et al. Cidar: Culturally relevant instruction dataset for arabic. *arXiv preprint arXiv:2402.03177*, 2024.

ArbML. Al jazeera news articles. https://huggingface.co/datasets/arbml/news_articles_aljazeera, 2023a. HuggingFace Dataset.

ArbML. Arsen-20. https://huggingface.co/datasets/arbml/Arsen_20, 2023b. HuggingFace Dataset.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning, 2024.

Francis Bond and Takayuki Kuribayashi. The Japanese Wordnet 2.0. In German Rigau, Francis Bond, and Alexandre Rademaker (eds.), *Proceedings of the 12th Global Wordnet Conference*, pp. 179–186, University of the Basque Country, Donostia - San Sebastian, Basque Country, January 2023. Global Wordnet Association. doi: 10.18653/v1/2023.gwc-1.22. URL https://aclanthology.org/2023.gwc-1.22/.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *arXiv preprint arXiv:2410.02677*, 2024.

Michael A Covington and Joe D McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100, 2010.

Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. Dsap: Analyzing bias through demographic comparison of datasets. *Information Fusion*, 115:102760, 2025.

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

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Md Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. Nativqa: Multilingual culturally-aligned natural query for llms. *arXiv preprint arXiv:2407.09823*, 2024.

Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. Vsm 2013 — values survey module 2013. https://geerthofstede.com/research-and-vsm/vsm-2013/, 2013. Accessed: 2025-01-04.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*, 2023.

Walaa Saber Ismail and Masun Nabhan Homsi. Dawqas: A dataset for arabic why question answering system. *Procedia computer science*, 142:123–131, 2018.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de Las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. corr, abs/2310.06825, 2023. doi: 10.48550. *arXiv preprint ARXIV.2310.06825*, 10, 2023.

Webdell Johnson. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15, 1944. URL https://pure.mpg.de/rest/items/item_2350946/component/file_2562008/content.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*, 2024.

Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 461–466, 2018.

"Fajri Koto, Haonan Li, Sara Shatanawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin". Arabicmmlu: Assessing massive multitask language understanding in arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Kunishou. Databricks dolly 15k (japanese). https://huggingface.co/datasets/kunishou/databricks-dolly-15k-ja, 2023. HuggingFace Dataset.

levelevel. Aozoratxt: Aozora bunko text corpus. https://github.com/levelevel/AozoraTxt, 2023. GitHub Repository.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216, 2024.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 110–119, 2016.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*, 2023.

Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pp. 281–297, 1967.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *arXiv preprint arXiv:2409.11404*, 2024.

msfm. Ichikara instruction all. https://huggingface.co/datasets/msfm/ichikara-instruction-all, 2023. HuggingFace Dataset.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*, 2023.

Jinlong Pang, Jiaheng Wei, Ankit Parag Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. Improving data efficiency via curating llm-driven rating systems. *arXiv preprint arXiv:2410.10877*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040/`.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pp. 1–96, 2025.

LLM-Book Project. Aio: Japanese instruction dataset. `https://huggingface.co/datasets/llm-book/aio`, 2023. HuggingFace Dataset; see also `https://sites.google.com/view/project-aio/dataset`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Cahtrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. Commonsense reasoning in arab culture, 2025.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*, 2023. URL `https://huggingface.co/datasets/FreedomIntelligence/EXAMs`.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning, 2024.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. Jaquad: Japanese question answering dataset for machine reading comprehension. *arXiv preprint arXiv:2202.01764*, 2022.

Masashi Takeshita and Kenji Araki. Jcommonsensemorality. In *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing (in Japanese)*, pp. 357–362, 2023.

Masashi Takeshita and Rafal Rzepka. Jethics: Japanese ethics understanding evaluation dataset, 2025. URL `https://arxiv.org/abs/2506.16187`.

Wikimedia Foundation. Wikimedia project dumps. `https://dumps.wikimedia.org/`, 2024a. Accessed 2024.

Wikimedia Foundation. Japanese wikinews. https://ja.wikinews.org/wiki/, 2024b. Accessed 2024.

Trong Wu. An accurate computation of the hypergeometric distribution function. *ACM Transactions on Mathematical Software (TOMS)*, 19(1):33–43, 1993.

Gregory Yauney, Emily Reif, and David Mimno. Data similarity is not enough to explain language model performance. *arXiv preprint arXiv:2311.09006*, 2023.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Tao Liu, and Deyi Xiong. Cmoraleval: A moral evaluation benchmark for chinese large language models, 2024. URL https://arxiv.org/abs/2408.09819.

Xinlu Zhang, Zhiyu Zoey Chen, Xi Ye, Xianjun Yang, Lichang Chen, William Yang Wang, and Linda Ruth Petzold. Unveiling the impact of coding data instruction fine-tuning on large language models reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25949–25957, 2025.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

A Limitations

While the correlations observed in our analysis frequently exceed $|0.40|$ and in many cases surpass $|0.70|$, indicating that linguistic structure is meaningfully associated with cultural-alignment performance, the number of finetuning datasets available per language (9–13) imposes constraints on statistical generality. Our findings should therefore be interpreted as *descriptive patterns* rather than definitive causal and statistical association claims. In addition, PCA-based analyses were conducted using a single random seed, which may introduce minor variability in the resulting principal components, although the same transformation was consistently applied across all evaluated subsets. Moreover, cultural-alignment performance was evaluated using pass@1 accuracy, reflecting a standardized single-response setting applied consistently across all models and datasets. While alternative decoding or aggregation strategies may yield different absolute scores, our analysis focuses on relative trends, which are less sensitive to this choice. Importantly, several trends replicate across three model families (LLaMA, Mistral, DeepSeek) and three languages (Arabic, Japanese, Chinese), indicating that the observed relationships reflect stable tendencies rather than sampling noise. Future work with larger dataset collections, more model architectures, multiple PCA initializations, and more comprehensive evaluation protocols would further strengthen the generality of these findings; however, conducting additional subsets and experiments was not feasible within our computational budget.

B Additional Related Work

Beyond the cultural alignment benchmarks discussed in the main paper, several adjacent lines of work study dataset properties from complementary perspectives. Prior analyses document cultural bias and representational gaps in large-scale corpora, showing that widely used sources such as Wikipedia are strongly Western-centric (Naous et al., 2023), and that Arabic post-training resources suffer from scarcity and imbalance (Alkhowaiter et al., 2025). Other work explores dataset curation strategies for cultural diversity, such as synthetic dialogue generation in CulturePark (Li et al., 2024), though these efforts primarily target cultural judgments or moderation rather than identifying linguistic dataset properties associated with alignment.

Outside the cultural alignment literature, dataset-centric analyses such as DSAP (Dominguez-Catena et al., 2025) examine demographic similarity across corpora, and broader surveys document systematic cultural misalignment in LLMs (Pawar et al., 2025). Related work on dataset quality and diversity investigates how data curation, filtering, and diversity-oriented selection strategies influence downstream model performance (Chen et al., 2024; Zhou et al., 2023; Pang et al., 2024). While informative, these studies do not directly connect dataset-level linguistic structure to downstream cultural performance, nor do they evaluate such properties using culture-specific benchmarks. Our work complements these efforts by focusing specifically on how measurable linguistic properties of fine-tuning datasets relate to cultural alignment outcomes across models and languages.

C Training Configuration

Table 2 summarizes the fine-tuning configuration used across all model families and datasets. These settings were kept consistent to ensure comparability across languages and experimental conditions.

Table 2: Training configuration used for all fine-tuning experiments.

Setting	**Value**
Model families	LLaMA, Mistral, DeepSeek
Fine-tuning method	QLoRA (4-bit quantization)
Batch size	8
Gradient accumulation steps	8
Learning rate	2×10^{-5}
Optimizer	AdamW
Warmup ratio	0.03
Max sequence length	2048
Training epochs	3
LoRA rank (r)	64
LoRA α	16
LoRA dropout	0.05
Hardware	2×A100 80GB

D Benchmarks by Language

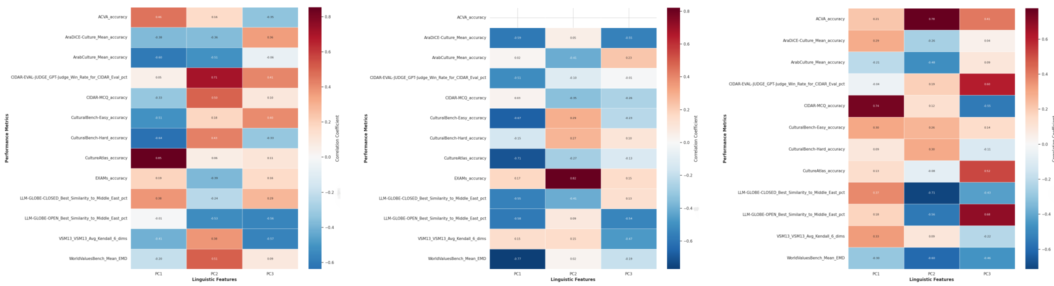
Table 3 lists the cultural and value-alignment benchmarks used for evaluation in each language. These benchmarks span multiple dimensions of cultural knowledge, norms, and moral reasoning.

Table 3: Cultural alignment benchmarks by category and language.

Language	**Cultural Knowledge**	**Cultural Values**	**Cultural Norms**
Arabic	CulturalBench (Easy, Hard) Chiu et al. (2024), CultureAtlas Fung et al. (2024), ArabCulture Sadallah et al. (2025), EXAMs-AR Sengupta et al. (2023)	VSM13 (Arabic) Hofstede et al. (2013), WorldValuesBench Zhao et al. (2024), ACVA-Arabic	CIDAR-MCQ, CIDAR-EVAL Alyafeai et al. (2024), AraDiCE Mousi et al. (2024), LLM-Globe (Open, Closed) Karinshak et al. (2024)
Japanese	CulturalBench (Easy, Hard) Chiu et al. (2024)	VSM13 Hofstede et al. (2013), WorldValuesBench Zhao et al. (2024)	J-Ethics Takeshita & Rzepka (2025), LLM-Globe (Open, Closed) Karinshak et al. (2024)
Chinese	CulturalBench (Easy, Hard) Chiu et al. (2024), CultureAtlas Fung et al. (2024), CEval-Exams Huang et al. (2023)	VSM13 Hofstede et al. (2013), WorldValuesBench Zhao et al. (2024)	CMoral Yu et al. (2024), LLM-Globe (Open, Closed) Karinshak et al. (2024)

Table 4: Explained variance ratios of the top three principal components for Arabic, Japanese, and Chinese datasets.

Language	PC1	PC2	PC3
Arabic	0.52	0.17	0.15
Japanese	0.52	0.34	0.10
Chinese	0.41	0.30	0.17



(a) AR–LLaMA (b) AR–Mistral (c) AR–DeepSeek

Figure 5: Correlation between dataset PCA components (PC1–PC3) and downstream cultural performance for Arabic across models.

E EXPLAINED VARIANCE RATIOS OF LANGUAGE-SPECIFIC PCA

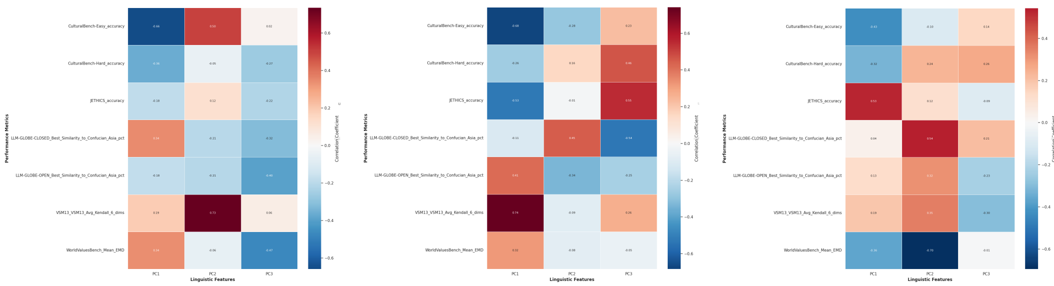
Table 4 reports the explained variance ratios of the top three principal components for each language. Across all languages, PC1 accounts for the largest share of variance (41–52

F ADDITIONAL CORRELATION HEATMAPS

This appendix presents the full correlation heatmaps between dataset PCA components (PC1–PC3) and downstream cultural benchmark performance, reported separately for each language and model. Each cell in Figs. 5, 6, and 7 shows the correlation coefficient between a PCA component score (columns) and a benchmark score (rows).

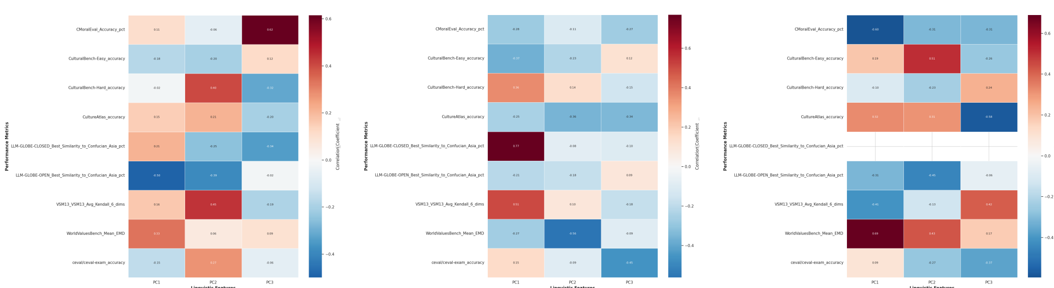
G THE USE OF ARTIFICIAL INTELLIGENCE

In the development of this paper, we employed artificial intelligence (AI) tools to enhance the quality of writing and ensure grammatical accuracy.



(a) JP–LLaMA (b) JP–Mistral (c) JP–DeepSeek

Figure 6: Correlation between dataset PCA components (PC1–PC3) and downstream cultural performance for Japanese across models.



(a) ZH–LLaMA (b) ZH–Mistral (c) ZH–DeepSeek

Figure 7: Correlation between dataset PCA components (PC1–PC3) and downstream cultural performance for Chinese across models.