An Experimental Study on the Influence of Culture on Cross-Lingual Sentiment Transfer

Ahao Liu, Chuanrong Wang, Haitong Yang*
Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning
Central China Normal University, Wuhan, Hubei 430079, PR China
The National Language Resources Monitor and Research Center for Network Media
Central China Normal University, Wuhan, Hubei 430079, PR China
{ahaoliu, rong806}@mails.ccnu.edu.cn, htyang@ccnu.edu.cn

Abstract

Identical linguistic expressions can convey different sentiments across cultural contexts. Yet, current multilingual models often reduce language to mere symbolic representation, neglecting the cultural pragmatics that fundamentally shape affective semantics in Sentiment Analysis (SA). Due to this oversight, the systematic performance degradation of Small Multilingual Language Models (SMLMs) on culturally distant targets is frequently attributed to resource constraints. This perspective obscures the pivotal role of cultural pragmatics, an intrinsic determinant of affective semantics, and thereby conceals cultural misalignment as the principal structural bottleneck. In this paper, we conduct a comprehensive empirical study to quantify the influence of culture on cross-lingual sentiment transfer across 7 common SMLMs and 5 linguistically diverse languages. By fitting a Linear Mixed-Effects Model (LMEM) (Baayen et al., 2008) on over 400 experimental runs, we disentangle cultural factors from confounding variables. Our results reveal that Cultural Distance is a significant, independent, negative predictor of transfer performance. Furthermore, representation probing and qualitative error analysis uncover a pragmatic alignment paradox: while SMLMs encode cultural distinctions, they fail to map these representations to downstream sentiment labels in high-context cultures (Hall, 1976). Ultimately, our work enhances the interpretability of cross-lingual transfer failures by statistically isolating cultural misalignment as a structural barrier, distinct from the resource constraints typically blamed for poor performance.

1 Introduction

The proliferation of SMLMs has revolutionized cross-lingual transfer, achieving remarkable success on benchmarks like XTREME (Hu et al., 2020), XGLUE (Liang et al., 2020), and Glot500 (ImaniGooghari et al., 2023). However, this success is not consistent across all languages or different cultural contexts. Empirical evidence suggests a systematic performance degradation when transferring from English to East Asian languages compared to Western European targets. This persistent asymmetry suggests that despite shared pretraining, current models struggle to bridge the gap between distinctly distant language groups. This evidence indicates that the multilingual landscape remains heavily skewed (Lauscher et al., 2020; Wu and Dredze, 2020), a disparity persisting even in modern architectures (Li et al., 2024).

Identifying the causes of this disparity remains a critical challenge. Prevailing explanations typically attribute transfer barriers to physical resource constraints (Wu and Dredze, 2020), surface-level representational mismatches (Muller et al., 2021; Wang et al., 2019), or deep structural divergences (Pires et al., 2019; Ahmad et al., 2019). While approaches relying on typological distances (Lin et al., 2019) or pragmatic proxies (Sun et al., 2021) attempt to mitigate these gaps, they overlook the social factors that are intrinsic to language use (Hovy and Yang, 2021). In particular, the distinct role of culture in tasks like SA remains underexplored (Hershcovich et al., 2022). Although sociolinguistic theories provide theoretical grounding (Hofstede, 2001; Hall, 1976), their quantitative impact on neural networks remains effectively entangled with the aforementioned linguistic and physical confounds.

In this work, we empirically evaluate the hypothesis that culture constitutes a distinct structural bottleneck for SMLMs, independent of linguistic and resource factors.

To scientifically quantify this effect, we conducted a large-scale evaluation across 7 common SMLMs (mBERT (Devlin et al., 2019), XLM-R (base and large) (Conneau et al., 2020), mT5 (Xue et al., 2021), mBART (Tang et al., 2020), Bloomz (Muennighoff et al., 2022) and Gemma-3-270M

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30936–30955

(Team et al., 2025)). Transcending observational correlations, we employ LMEM (Baayen et al., 2008) to perform rigorous statistical attribution. By constructing a hierarchical feature set that contains the physical-environmental (Geographic Distance, Resource Scale), the symbolic surface (Orthographic Distance (Littell et al., 2017), OOV Rate) and the linguistic kernel (Syntactic Distance (Littell et al., 2017), Emotion Distance (Sun et al., 2021)) layers, we statistically isolate the independent impact of culture. Our analysis reveals that Cultural Distance remains a robust, negative predictor of transfer performance even after strictly controlling for confounding factors. Moreover, the persistent performance gap between Western and Eastern targets is not merely an artifact of resource scarcity or syntactic divergence, but stems from a structural failure to align high-context pragmatic representations. Our observations can be summarized as follows:

First, utilizing LMEM, we statistically disentangle Cultural Distance from linguistic confounds, identifying it as a significant, negative bottleneck for transfer performance. Second, our representation probing reveals a pragmatic alignment paradox: while models successfully encode cultural distinctions, they fail to align these disjoint subspaces for downstream inference. Third, these findings underscore that for pragmatically charged tasks, prioritizing cultural homophily is as critical as linguistic similarity in source language selection.

2 Related Work

2.1 Linguistic and Resource Adaptation Mechanisms

Recent research on adapting SMLMs to low-resource languages has largely targeted physical and surface barriers. Strategies range from data augmentation (MulDA (Liu et al., 2021)) to vocabulary-centric methods like dictionary transfer (Sakajo et al., 2025) and trans-tokenization (Remy et al., 2024), aiming to mitigate corpus scarcity. In the realm of parameter efficiency, the field has shifted toward modular approaches rooted in LoRA (Hu et al., 2022). While these adapters prove effective in multilingual summarization (Whitehouse et al., 2024) and translation (Sel and Hanbay, 2024), recent studies highlight their limitations in handling unseen languages (Schlenker et al., 2025) and bridging few-shot gaps via contrastive learning (Borchert et al., 2025). Crucially, while these mechanisms enhance structural adaptation (syntax and lexicon), they operate primarily on explicit linguistic forms, often neglecting the implicit pragmatic shifts induced by cultural misalignment.

2.2 Semantic Alignment in Sentiment Analysis

Beyond surface translation, cross-lingual SA has evolved from traditional contrastive alignment (Lin et al., 2023) to generative and structural modeling. Recent approaches employ data augmentation based on large language models (LLMs) (Šmíd et al., 2025) or incorporate richer contexts via hybrid relational graphs (Ji et al., 2025) and multi-scale optimization (Wu et al., 2025). However, these techniques largely target explicit semantic equivalence. Comprehensive evaluations indicate that while models capture literal sentiment well (Zhang et al., 2024), they fail to decode high-context nuances such as indirectness or politeness strategies where sentiment polarity is culturally conditional rather than lexically defined.

2.3 Computational Modeling of Culture

The integration of culture into natural language processing (NLP) has shifted from passive prediction to active simulation, grounded in emerging sociocultural frameworks (Zhou et al., 2025). Empirically, research now spans data optimization for cultural representativeness (Yao et al., 2025), evolutionary simulation via LLMs agents (Perez et al., 2024; Vallinder and Hughes, 2024), and the quantification of Western bias through benchmarks like CultureLLM (Li et al., 2024).

However, a critical methodological gap persists: rigorous statistical disentanglement. Most prior works are either observational (probing bias) or generative (simulating dynamics). Few have rigorously disentangled the performance degradation strictly attributable to Cultural Distance from entangled linguistic or resource confounds. Our work fills this gap by employing LMEM to statistically isolate the independent impact of cultural misalignment in downstream tasks.

3 Method

In this work, we propose a statistical framework to quantify the impact of cultural misalignment on cross-lingual transfer. We first operationalize cultural and linguistic distances into computable metrics, and then introduce a LMEM to disentangle the independent contribution of culture from

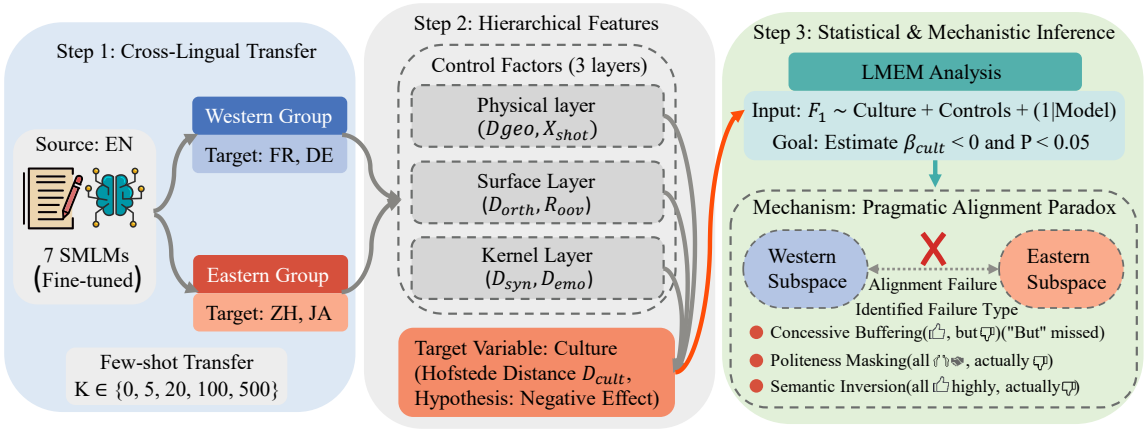

Figure 1: Overview of our experimental and analytical framework. **(Step 1)** We evaluate zero-shot and few-shot transfer performance. **(Step 2)** We construct hierarchical feature sets. **(Step 3)** These are integrated into a LMEM to identify cultural effects and analyze the mechanism underlying the pragmatic alignment paradox.

confounding factors. The overall architecture of our proposed framework is illustrated in Figure 1.

3.1 Quantifying Cultural Distance

To quantify the cultural divergence between the source language (L_S) and the target language (L_T), we adopt Hofstede's Cultural Dimensions Theory[1] (Hofstede, 2001). This framework characterizes culture through six dimensions: Power Distance (PDI), Individualism (IDV), Masculinity (MAS), Uncertainty Avoidance (UAI), Long Term Orientation (LTO), and Indulgence (IVR).

We map each language to its representative nation (e.g., English → UK, Chinese → China) to retrieve the corresponding dimensional scores. Let vector $\mathbf{v}_L \in R^6$ represent the normalized cultural profile of language L. The Cultural Distance (D_{cult}) is calculated as the Euclidean distance between the source and target vectors:

$$D_{cult}(L_S, L_T) = \sqrt{\sum_{k=1}^{6}(v_{S,k} - v_{T,k})^2} \quad (1)$$

where $v_{S,k}$ and $v_{T,k}$ denote the score of the k-th dimension for the source and target languages, respectively. Table 1 presents the cultural profiles for the languages studied in this work.

3.2 Hierarchical Control Features

To isolate the independent impact of culture and mitigate confounding effects from correlated linguistic and physical factors, we construct a hierarchical set of controllable variables. We acknowledge that Cultural Distance often correlates with other proxies, necessitating careful attribution analysis. The variables are categorized into three layers:

[1]Details of Hofstede's Cultural Dimensions and method of language-to-country mapping are provided in §A.

Language	PDI	IDV	MAS	UAI	LTO	IVR
English (EN)	35	89	66	35	51	69
German (DE)	35	67	66	65	83	40
French (FR)	68	71	43	86	63	48
Chinese (ZH)	80	20	66	30	87	24
Japanese (JA)	54	46	95	92	88	42

Table 1: Hofstede's Cultural Dimension scores for the source (English) and target languages. Note the distinct divergence between Western (EN, DE, FR) and Eastern (ZH, JA) clusters.

Physical-Environmental Layer

- **Geographic Distance** (D_{geo}): Calculated as the great-circle distance between languages' representative coordinates (URIEL (Littell et al., 2017)), used as a proxy for spatial proximity and potential data-sourcing bias.
- **Resource Scale** (X_{shot}): Quantified by the logarithmic transformation of the number of few-shot training examples ($\log(1 + N_{shot})$), controlling for the impact of data availability.

Surface-Symbolic Layer

- **OOV Rate** (R_{oov}): Measured as the percentage of tokens in the target test set that are fragmented into sub-optimal subwords or OOV tokens by the source model's tokenizer. This serves as a proxy for tokenization fragmentation and subword fertility issues.

- **Orthographic Distance** (D_{orth}): Quantified by the edit distance between the orthographic scripts (e.g., Latin vs. CJK characters) of the source and target languages, adapted from URIEL (Littell et al., 2017).

Linguistic-Kernel Layer

- **Syntactic Distance** (D_{syn}): Calculated based on the cosine distance of syntactic dependency feature vectors (URIEL (Littell et al., 2017)), controlling for structural linguistic differences.

- **Emotion Distance** (D_{emo}): We calculate D_{emo} to measure lexical similarities of emotions across languages. Based on the Emotion Semantics Database (ESD) (Jackson et al., 2019), this metric captures the structural divergence of emotion concept networks between the source and target languages[2].

All continuous variables, including the target predictor **Cultural Distance** (D_{cult}), are Z-score normalized prior to regression analysis. This ensures that all coefficients are on a comparable scale, reflecting effect size. Table 2 provides a concise summary of all operationalized symbols, metric definitions, and their corresponding data sources[3].

3.3 Statistical Framework

To rigorously disentangle the independent impact of cultural factors while controlling for other influential variables, we employ LMEM (Baayen et al., 2008). LMEM is suitable for our data structure, which exhibits nesting (multiple shots within languages, nested within models), thus accounting for model-specific random effects and avoiding spurious correlations.

We specify the LMEM as follows, with the F1 score as the dependent variable:

$$F1_{ij} = \beta_0 + \beta_1 X_{\text{shot}} + \beta_2 D_{\text{syn}} + \beta_3 D_{\text{emo}} + \beta_4 D_{\text{cult}} + u_i + \epsilon_{ij} \quad (2)$$

where $F1_{ij}$ denotes the performance of Model i on Language j, and β_0 is the global intercept. X_{shot} represents the log-transformed Resource Scale. The predictors D_{syn}, D_{emo}, and D_{cult} correspond to the standardized Syntactic Distance, Emotion Distance, and Cultural Distance, respectively. The coefficients $\beta_1, \beta_2, \beta_3$, and $\boldsymbol{\beta_4}$ quantify the fixed effects, with our primary interest lying in $\boldsymbol{\beta_4}$ (cultural influence). Finally, the term $u_i \sim \mathcal{N}(0, \sigma_u^2)$ accounts for the model-specific random intercept, and ϵ_{ij} denotes the residual error.

Statistical Evaluation. Significance is assessed via Wald z-tests ($p < 0.05$) and model fit via Akaike Information Criterion (AIC), reporting standardized coefficients (β) for effect size. Notably, variables exhibiting multicollinearity are excluded from Eq. 2 to ensure stability and detailed diagnostics are provided in §5.

4 Experiment

To empirically verify the cultural impact hypothesis, we conduct a large-scale evaluation involving over 400 experimental runs. This section details the data curation, model configurations, and the transfer protocol.

4.1 Datasets and Curation

We utilize the Multilingual Amazon Reviews Corpus (MARC) (Keung et al., 2020), a large-scale collection of customer reviews. The original dataset contains 200,000 training samples, 5,000 validation and test samples for each of the six languages[4].

To construct a culturally balanced benchmark, we perform two curation steps:

- **Cultural Clustering:** We select 5 languages to form two distinct cultural clusters: the Western Group (EN, FR, DE) and the Eastern Group (ZH, JA). We explicitly exclude Spanish from the Western group[5].

- **Data Downsampling:** To simulate realistic low resource transfer scenarios, we downsample the training set to 20,000 balanced samples per language using a fixed random seed (seed = 42). The validation and test sets are kept at their original size (5,000 samples) to ensure stable evaluation.

4.2 Models and Transfer Protocol

We evaluate 7 common SMLMs spanning three distinct architectures to ensure the generalizability

[2]The explicit mathematical formulation for this metric is detailed in §F.

[3]The precise per-language statistics for all evaluated metrics are detailed in §G.

[4]The original MARC dataset includes six languages: English, French, German, Spanish, Chinese, and Japanese.

[5]Spanish was excluded to maintain linguistic balance other than English and to prevent phylogenetic bias in statistical attribution.

Symbol	Variable Metric	Source / Method
Dependent Variable		
$F1_{ij}$	Sentiment F1 Score	Experimental Results
Target Predictor		
D_{cult}	Cultural Distance	Hofstede Dim (Hofstede, 2001)
Control Layers (Confounders)		
D_{syn}	Syntactic Distance	URIEL Dependency (Littell et al., 2017)
D_{emo}	Emotion Distance	ESD (Jackson et al., 2019)
D_{orth}	Orthographic Distance	URIEL Script (Littell et al., 2017)
R_{oov}	OOV Rate	Tokenizer Fertility
D_{geo}	Geographic Distance	URIEL Centroids (Littell et al., 2017)
X_{shot}	Resource Scale	$\log(1 + \text{Shot})$

Table 2: Summary of operationalized variables and data sources used in our statistical attribution framework. The variables are categorized into hierarchical control layers as defined in §3.2.

Arch	Model Name	# of Param
Enc.	mBERT	110M
Enc.	XLM-R-base	250M
Enc.	XLM-R-large	560M
Enc-Dec.	mT5-base	580M
Enc-Dec.	mBART-large-50	610M
Dec.	Bloomz	560M
Dec.	Gemma-3	270M

Table 3: Overview of the 7 SMLMs evaluated, categorized by their architecture and parameter counts.

of our findings. The specifications for each model are summarized in Table 3.

Two-Stage Transfer Framework. To rigorously assess transferability while mitigating catastrophic forgetting, we implement a unified two-stage protocol:

- **Source Fine-tuning (English Prior):** We first fine-tune all models on the full English training set (20k samples). This involves either a classification head fine-tuning for encoder-only models or instruction-tuned generation for encoder-decoder and decoder-only models, establishing a strong English Culture Prior.
- **Target Adaptation via LoRA:** In the cross-lingual phase, we freeze the pre-trained backbones and inject LoRA (Hu et al., 2022) modules. We adapt the English-tuned models to target languages under 5 few-shot settings ($k \in \{0, 5, 20, 100, 500\}$).

Detailed implementation are provided in §B.

4.3 Statistical Data Preparation

To enable the LMEM analysis described in §3, we construct a hierarchical analysis dataset. We aggregate the F1 score from all experimental runs (7 models $\times$ 4 target languages $\times$ 5 shots) as the dependent variable ($F1_{ij}$). These performance metrics are then aligned with the pre-calculated feature vectors (e.g., Cultural Distance, Syntactic Distance, and Emotion Distance) for each specific language pair. This consolidated dataset serves as the input for our statistical attribution analysis in the subsequent section.

5 Quantitative Analysis

In this section, we empirically verify the independent influence of Cultural Distance. We first present the macroscopic performance landscape, followed by a rigorous attribution analysis using LMEM, and finally quantify the effect size.

5.1 Performance Landscape

Figure 2 illustrates the few-shot transfer performance (F1) of 7 SMLMs from the source language (English) to four target languages. We observe two consistent patterns:

- **Asymptotic Growth:** Increasing samples from 0 to 500 yields performance gains across all languages.

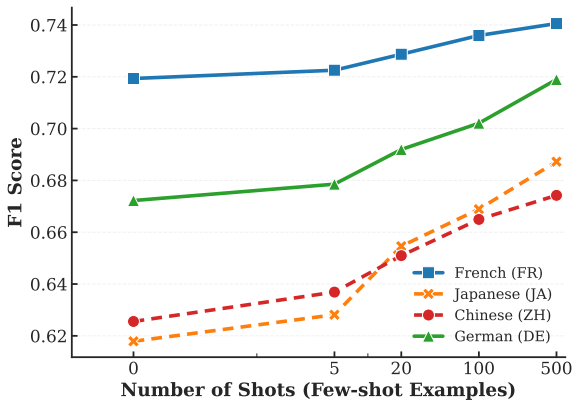

Figure 2: Cross-lingual transfer performance (F1) across 7 SMLMs. A persistent performance gap remains between culturally close (FR, DE) and distant (ZH, JA) targets even as training shots increase.

ID	Controls Included	β_{cult}	P-val	Status
Level 1: Baseline				
M1	+ X_{shot}	-0.024	**0.000**	Significant
Level 2: Physical & Environmental				
M2	+ D_{geo}	+0.003	0.746	Collinear
Level 3: Surface-Symbolic Proxies				
M4	+ D_{orth}	-0.003	0.578	Collinear
M22	+ R_{oov}	+0.026	**0.020**	Collinear
Level 4: Linguistic Kernel Isolation				
M14	+ D_{syn}	-0.008	0.347	Masked
M7	**+ D_{syn} + D_{emo}**	**-0.016**	**0.046**	**Robust (-)**

Table 4: Fixed effects of Cultural Distance (D_{cult}) across four hierarchical levels. While physical (M2) and surface proxies (M4, M22) absorb explanatory power due to collinearity, the cultural impact remains robustly negative (M7) after controlling for linguistic kernel factors.

- **The Cultural Cliff:** Crucially, a distinct performance gap persists between the Western Group (FR, DE) and Eastern Group (ZH, JA). Even at 500-shot, the trajectories remain parallel, suggesting that mere data scaling cannot bridge the underlying structural divergence.

5.2 Attribution Analysis via Hierarchical LMEM

To isolate the primary structural drivers of this gap, we conducted a step-wise regression analysis using LMEM. We specified the model structure as defined in Eq. (2), systematically introducing control variables to isolate the independent contribution of **Cultural Distance** (D_{cult})[6]. Table 4 summarizes the key findings across different feature levels[7].

The Baseline Reality. Initially, in the baseline model M1 controlling only for shot size, Cultural Distance exhibits a strong, statistically significant, negative correlation with transfer performance ($\beta = -0.024, p < 0.001$). This confirms the macroscopic observation that cultural divergence is a primary predictor of performance drops.

Proxies and Collinearity. We subsequently introduce controls from the physical (D_{geo}, M2) and surface-symbolic (D_{orth}, M4; R_{oov}, M22) layers. In all three specifications, the coefficient for Cultural Distance loses significance or exhibits artifactual positivity. This stems from the high multicollinearity inherent in our typological split: geographic isolation often historically drives cultural divergence, while distinct cultural clusters (e.g., East Asian) utilize distinct writing systems (CJK vs. Latin), manifesting as high Orthographic Distance and OOV Rate. Consequently, these tangible metrics act as strong statistical proxies, absorbing the explanatory power of the latent cultural variable. This indicates that while culture is the root cause, its impact is partially mediated through and statistically masked by these physical separation and tokenization barriers. Crucially, the significance of Model M7 proves that cultural misalignment imposes a penalty that cannot be resolved by merely fixing tokenization (i.e., lowering OOV Rate).

Deep Pragmatic Independence. The critical insight emerges in **Model M7**, where we control for the linguistic kernel Syntactic Distance[8] and Emotion Distance. Unlike surface features, these controls disentangle structural and lexical semantics from cultural pragmatics. Crucially, as can be seen from Figure 3, the significance of Cultural Distance is restored with a negative coefficient ($\beta = -0.016, p < 0.05$). This reveals a statistical **suppression effect**: accounting for structural and emotional variance unmasks the latent cultural signal. This validates that cultural misalignment functions as a distinct, high-level pragmatic bottleneck, persisting even when syntactic and token-level barriers are bridged.

[6]To address potential concerns regarding the statistical power given the limited number of target languages, we conducted a post-hoc Monte Carlo power analysis. Detailed simulation setups and results are provided in §C.

[7]Due to space limitations, the complete regression results and analysis for combinations are provided in §C.

[8]The positive β reflects Transformers' robustness to syntactic reordering (via self-attention), rendering structural di-

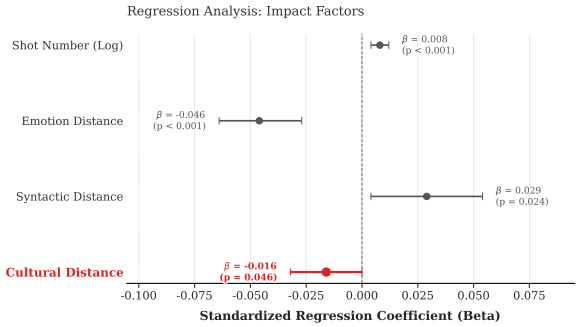

Figure 3: Forest plot of standardized coefficients for **Model M7**. The error bars represent 95% confidence intervals. Note that Cultural Distance remains significantly negative even after controlling for Syntactic Distance and Emotion Distance.

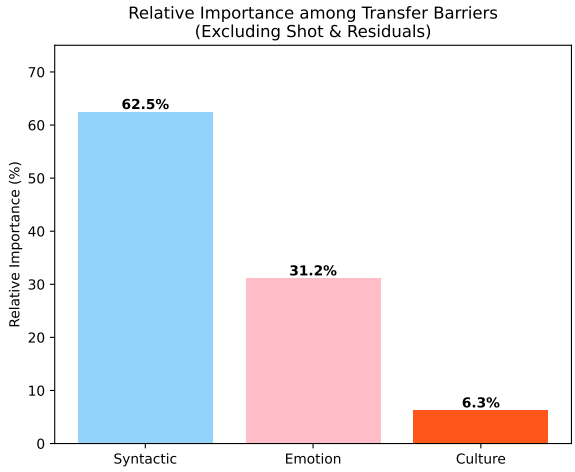

Figure 4: Variance partitioning of the F1 score. Cultural Distance contributes a unique explanatory power (6.3%) independent of linguistic factors.

5.3 Variance Partitioning

To quantify the magnitude of this effect, we performed variance partitioning analysis based on the robust M7 specification. As shown in Figure 4, Cultural Distance uniquely explains approximately **6.3%** of the model's performance variance.

While Syntactic Distance remains the dominant factor, the independent contribution of culture is non-negligible. Crucially, this 6.3% represents the pragmatic gap, the portion of performance loss that cannot be recovered by simply improving tokenizers or adding syntactic parsers, highlighting the necessity for culturally-aware alignment strategies.

6 Mechanism and Qualitative Analysis

While the statistical results confirm the independent negative impact of Cultural Distance, questions remain regarding the internal mechanism of this failure. Does the model fail to encode cultural distinctions, or does it fail to align them for the downstream task? To answer this, we conduct representation probing and a fine-grained qualitative error analysis using the zero-shot transfer models.

6.1 The Pragmatic Alignment Paradox

We investigate the geometry of the model's latent space by training a linear diagnostic classifier on the frozen representations (e.g., [CLS] embeddings for mBERT) to distinguish between Western (FR, DE) and Eastern (ZH, JA) samples.

Perfect Separability vs. Transfer Failure. As visualized in Figure 5, the t-SNE projection reveals a stark geometric reality[9]: the embedding spaces of Western and Eastern languages are structurally disjoint, forming two isolated clusters with no overlap. Quantitatively, the probe achieves a classification accuracy $> 99\%$ across all architectures. This result presents a critical **Pragmatic Alignment Paradox:**

- **High Encoding Capacity:** The model is not culturally blind. It perfectly distinguishes cultural domains, likely driven by script and lexical patterns.
- **Alignment Failure:** Despite this clear distinction, the sentiment classifier fine-tuned on the Western cluster (EN) learns decision boundaries that are geometrically valid only within that local subspace. When applied to the disjoint Eastern subspace, these boundaries fail to generalize.

Thus, the cultural performance gap is not due to a lack of representational capacity, but rather a failure of cross-cultural alignment. The model knows the culture is different but lacks the bridge to map Eastern pragmatic patterns to Western-defined sentiment labels.

6.2 Qualitative Error Analysis

To understand how this alignment failure manifests in actual inference, we manually inspected

vergence less critical than Cultural Distance.

[9]The t-SNE diagrams for the remaining architectures and further representation probing results are shown in §D.

ID	Original Text	English Translation	Gold	Pred	Error Type
1	书的质量很好，卡片很好，唯一的不足就是光盘读不了。	The book and cards are of high quality. **The only downside is** that the CD is unreadable.	Neg	Pos	**Concessive Buffering**
2	亲尽快退款，谢谢。我希望你们尽快公平公正的处理。	**Dear**, **please** refund ASAP, **thank you**. I **hope** you handle this **fairly**.	Neg	Pos	**Politeness Masking**
3	有种被骗的感觉，通关笔记是一本本子。卖家真会做生意！	Feel cheated, The guide is just an empty notebook. The seller **really knows how to do business!**	Neg	Pos	**Sarcasm**
4	気に入ってるのですが、弱いです。部品の予備が欲しいです。	I **really like** it, but it's fragile. I'd like to get some spare parts.	Neg	Pos	**Concessive Buffering**
5	まだ、荷物が届かないのですがどのような状況でしょうか？連絡お待ちしてます。	My package hasn't arrived yet. **Could you please** tell me the current status? I **look forward to** hearing from you.	Neg	Pos	**Politeness Masking**

Table 5: Representative error cases in zero-shot transfer to high-context cultures (JA, ZH). The models consistently fail to decode pragmatic cues like indirect hedging, politeness markers masking complaints, and sarcasm, leading to polarity misclassification.

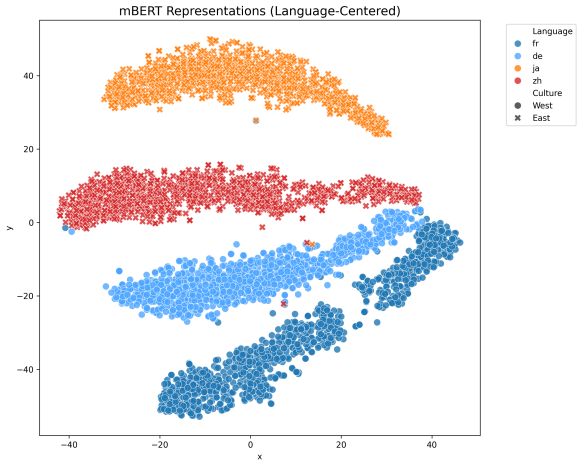

Figure 5: t-SNE visualization of mBERT representations. The embedding space exhibits a severe geometric split between Western and Eastern clusters, explaining why the classifier head fails to transfer effectively.

error cases in the Eastern target languages (ZH, JA). We identified consistent patterns of Pragmatic Failures, where the model interprets literal semantics correctly but misses the culturally embedded illocutionary force. Table 5 presents representative examples[10].

Failure Type 1: Concessive Buffering. In Eastern cultures, negative feedback is often prefaced by elaborate praise to maintain social harmony. As shown in **Case 1**, SMLMs fine-tuned on English tend to over-attend to the initial positive lexical cues. Consequently, the models fail to recognize the adversative shift, allowing the buffer content to dominate the sentiment prediction, resulting in false positives.

Failure Type 2: Politeness Masking Negativity. A pervasive error source stems from the misalignment between social stance (politeness) and sentiment polarity. As illustrated in **Case 2**, Japanese and Chinese reviews frequently employ honorifics or gratitude expressions even when lodging a complaint. The models conflate these politeness markers with positive sentiment, failing to decouple the polite form of the message from its dissatisfied content.

Failure Type 3: Sarcasm and Semantic Inversion. The third category involves semantic inversion, where the surface meaning contradicts the intended illocutionary force. In **Case 3**, the model interprets hyperbolic praise literally, missing the pragmatic implicature triggered by contextual incongruities. This underscores the model's inability to detect sentiment when it relies on deep cultural decoding rather than explicit lexical markers.

7 Conclusion

In this work, we present the first rigorous attribution analysis quantifying the unique explanatory power of Cultural Distance on cross-lingual sentiment transfer. By conducting a large-scale evaluation across 7 SMLMs and 5 diverse languages, employing hierarchical LMEM, we systematically

[10] A systematic inspection reveals that these specific pragmatic failures stubbornly persist even under the 500-shot regime. See E.

disentangled cultural factors from physical, surface-symbolic, and structural confounds. We demonstrate that while surface-symbolic and linguistic-kernel features mediate part of the impact, Cultural Distance remains a robust, negative predictor ($\beta < 0, p < 0.05$), underscoring that shared syntax does not guarantee shared pragmatics. Furthermore, our mechanistic analysis uncovers a Pragmatic Alignment Paradox: models possess a high capacity to encode cultural distinctions but fail to align these representations for downstream tasks due to the geometric disjointedness of cultural subspaces. Consequently, we argue that the field must advance beyond **Linguistic Alignment** toward **Cultural Alignment**, developing culturally-aware fine-tuning strategies to bridge the deep semantic gap that brute-force data scaling alone cannot resolve.

Limitations

While our study offers robust statistical evidence, we acknowledge several limitations. First, our analysis is restricted to SMLMs. It remains an open question whether the emergent cultural reasoning capabilities of LLMs can naturally mitigate the observed Pragmatic Alignment Paradox, or if they merely amplify distinct cultural biases. Second, relying on Hofstede's static indices acts as a high-level proxy, which may oversimplify intra-cultural nuances and dynamic norm shifts, potentially risking cultural essentialism. Third, while our language selection maximizes the Western-Eastern contrast, the exclusion of Global South languages (e.g., Afro-Asiatic languages) limits universality. Furthermore, our findings are anchored in SA, a task inherently laden with pragmatics; the extent to which Cultural Distance impacts fact-centric tasks (e.g., Entity Extraction) remains to be verified. Finally, this work focuses on diagnostic attribution rather than algorithmic mitigation. We rigorously identify Cultural Distance as a distinct bottleneck but do not propose specific interventions (e.g., pragmatic adapters), leaving the development of solution-oriented strategies to future research.

Ethical Considerations

A primary ethical consideration in this work is the operationalization of culture via Hofstede's Cultural dimensions. We explicitly caution against an essentialist interpretation of these metrics. The utilized cultural scores are aggregate statistical proxies intended for macroscopic computational analysis and do not reflect the diversity, agency, or dynamic nature of individuals within those linguistic groups. Our methodological clustering of languages into Western and Eastern groups is a controlled experimental setup to isolate variance, not a prescriptive sociopolitical categorization. Future research should strive for more granular, non-binary cultural representations to avoid reinforcing stereotypes.

Our research is motivated by the observation that current multilingual models disproportionately underperform on high-context Asian languages, creating a technological equity gap. By diagnosing the root cause as a Pragmatic Alignment Paradox rather than mere resource scarcity, we aim to direct community attention toward culturally inclusive alignment strategies, fostering fairer NLP systems for diverse global populations.

We utilize the publicly available MARC, adhering to its usage licenses. Our focus on SMLMs ensures that our large-scale attribution analysis maintains a relatively low carbon footprint compared to training or fine-tuning LLMs.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. Furthermore, the authors express their gratitude to Tingting He for her constructive feedback and insightful discussions, and to Xuhui Tian, Xiaomeng Wang, and Zaibin Duan for their assistance with preliminary data processing.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 2440–2452.

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Philipp Borchert, Jochen De Weerdt, and Marie-Francine Moens. 2025. Bridging language gaps: Enhancing few-shot language adaptation. *arXiv preprint arXiv:2508.19464*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Edward T Hall. 1976. *Beyond culture*. Anchor.

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

G. Hofstede. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. SAGE Publications.

Geert H. Hofstede, Gert Jan. Hofstede, and Michael. Minkov. 2010. *Cultures and organizations : software of the mind : intercultural cooperation and its importance for survival*, 3rd ed. edition. McGraw-Hill, New York.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. page 1082–1117.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Hongru Ji, Xianghua Li, Mingxin Li, Meng Zhao, and Chao Gao. 2025. Hybrid relational graphs with sentiment-laden semantic alignment for multimodal emotion recognition in conversation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 2973–2981.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A Smith. 2020. The multilingual amazon reviews corpus. *arXiv preprint arXiv:2010.02573*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and 1 others. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Nankai Lin, Yingwen Fu, Xiaotian Lin, Dong Zhou, Aimin Yang, and Shengyi Jiang. 2023. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2935–2946.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, and 1 others. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. *arXiv preprint arXiv:2101.11109*.

Jérémy Perez, Corentin Léger, Marcela Ovando-Tellez, Chris Foulon, Joan Dussauld, Pierre-Yves Oudeyer, and Clément Moulin-Frier. 2024. Cultural evolution in populations of large language models. *arXiv preprint arXiv:2403.08882*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *arXiv preprint arXiv:2408.04303*.

Haruki Sakajo, Yusuke Ide, Justin Vasselli, Yusuke Sakai, Yingtao Tian, Hidetaka Kamigaito, and Taro Watanabe. 2025. Dictionaries to the rescue: Cross-lingual vocabulary transfer for low-resource languages using bilingual dictionaries. *arXiv preprint arXiv:2506.01535*.

Julian Schlenker, Jenny Kunz, Tatiana Anikina, Günter Neumann, and Simon Ostermann. 2025. Only for the unseen languages, say the llamas: On the efficacy of language adapters for cross-lingual transfer in english-centric llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 849–871.

Ilhami Sel and Davut Hanbay. 2024. Efficient adaptation: Enhancing multilingual models for low-resource language translation. *Mathematics*, 12(19):3149.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2025. Laca: Improving cross-lingual aspect-based sentiment analysis with llm data augmentation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–853.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Aron Vallinder and Edward Hughes. 2024. Cultural evolution of cooperation among llm agents. *arXiv preprint arXiv:2412.10270*.

Zihan Wang, Stephen Mayhew, Dan Roth, and 1 others. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1202–1228.

Chengyan Wu, Bolei Ma, Ningyuan Deng, Yanqing He, and Yun Xue. 2025. Multi-scale and multi-objective optimization for cross-lingual aspect-based sentiment analysis. *arXiv preprint arXiv:2502.13718*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.

Jing Yao, Xiaoyuan Yi, Jindong Wang, Zhicheng Dou, and Xing Xie. 2025. Caredio: Cultural alignment of llm via representativeness and distinctiveness guided data optimization. *arXiv preprint arXiv:2504.08820*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

A Operationalization of Hofstede's Cultural Dimensions

To rigorously quantify the Cultural Distance (D_{cult}) between source and target languages, we adopt the six-dimensional framework established by Hofstede (2001) and refined in later works (Hofstede et al., 2010). While originally developed for cross-cultural psychology, these dimensions provide a robust, vectorizable schema for capturing

the latent social norms that govern pragmatic communication. In our experimental setup, we map each language to the cultural scores of its representative nation (e.g., English → UK, Chinese → China). Below, we detail the six dimensions and their specific implications for cross-lingual SA.

A.1 The Six Dimensions and SA Implications

Power Distance Index (PDI). PDI measures the extent to which less powerful members of a society accept and expect that power is distributed unequally.

- **Relevance to SA:** In high PDI cultures (e.g., ZH), communication tends to be more formal and hierarchical. Users may employ honorifics or indirect language when criticizing entities perceived as having higher status (e.g., established brands or sellers), potentially masking negative sentiment with polite surface forms. In contrast, low-PDI cultures (e.g., EN, DE) favor directness and egalitarian expression, leading to explicit sentiment labeling.

Individualism vs. Collectivism (IDV). This dimension is the most significant differentiator between our Western and Eastern clusters. It measures the degree of interdependence a society maintains among its members.

- **Relevance to SA:** High IDV cultures (Western Group) encourage explicit, "I"-centered expression of personal opinion (e.g., *"I absolutely hate this product"*). Low-IDV or Collectivist cultures (Eastern Group) prioritize social harmony and face-saving (*mianzi*). Negative feedback is often mitigated by **Concessive Buffering** prefacing criticism with praise to avoid disrupting social harmony which frequently confuses models trained on explicit Western data into predicting positive labels for negative reviews.

Uncertainty Avoidance Index (UAI). UAI defines the extent to which members of a culture feel threatened by ambiguous or unknown situations.

- **Relevance to SA:** High UAI cultures (e.g., JA, FR) tend to have rigid codes of belief and behavior. In product reviews, this may manifest as a preference for detailed, structured feedback and a lower tolerance for product defects. Linguistically, high UAI correlates with the usage of hedging devices (e.g., *"it might be the case that..."*) to mitigate the risk of making absolute claims, adding noise to the sentiment signal that Western-centric models may misinterpret as neutrality.

Masculinity vs. Femininity (MAS). This dimension focuses on the motivation of values: wanting to be the best (Masculine) vs. liking what you do (Feminine).

- **Relevance to SA:** High MAS cultures (e.g., JA, EN) value achievement and success. Reviewers may focus heavily on performance metrics, efficiency, and status. Low-MAS cultures focus more on quality of life and consensus. While less structurally disruptive than IDV, MAS divergence contributes to the lexical drift in what features are considered "positive" (e.g., aggressive performance vs. user comfort).

Long-Term Orientation (LTO). LTO describes how a society maintains links with its own past while dealing with the challenges of the present and future.

- **Relevance to SA:** High LTO societies (e.g., ZH, JA) view pragmatic adaptation as a virtue. In the context of e-commerce, this often correlates with patience regarding logistics or a focus on long-term product durability rather than immediate gratification. This shift in temporal focus can alter the semantic weight of time-related keywords in sentiment classification.

Indulgence vs. Restraint (IVR). IVR measures the extent to which people try to control their desires and impulses.

- **Relevance to SA:** High Indulgence cultures (e.g., EN) allow relatively free gratification, leading to uninhibited and polarized sentiment expression (strongly positive or negative). High-Restraint cultures (e.g., ZH) regulate gratification through strict social norms. This often leads to **Semantic Flattening**, where strong emotions are understated or encoded via subtle context rather than strong emotional adjectives, causing classifiers to underestimate the intensity of sentiment.

A.2 Language-to-Nation Mapping Rationale

To obtain concrete vector representations $\mathbf{v}_L \in R^6$, we follow standard conventions in cultural NLP (Lin et al., 2019; Hershcovich et al., 2022) by mapping languages to their primary linguistic dominance or phylogenic origin.

- **English (EN) → United Kingdom:** Selected as the phylogenic root. While US scores are comparable, UK scores minimize variance when paired with European neighbors like France and Germany in the Western cluster. Experimental verification confirms that mapping to the US still yields statistically significant conclusions.
- **French (FR) → France:** Selected as the primary center of the Francophone world.
- **German (DE) → Germany:** Selected as the linguistic epicenter of the Germanic cluster.
- **Chinese (ZH) → China:** Representing the dominant Sinophone culture and the source of the Mandarin dialect used in the corpus. This mapping strictly aligns with the dataset provenance, as the Chinese reviews in MARC are exclusively sourced from Amazon.cn (Mainland China).
- **Japanese (JA) → Japan:** A mono-cultural linguistic isolate.

We acknowledge that languages are spoken across diverse nations (e.g., French in Canada, German in Switzerland). However, our selection of specific target nations aims to maximize the distinct cultural signals associated with these primary nation-states, thereby increasing the statistical power of our attribution analysis while maintaining experimental tractability.

B Detailed Implementation Specifications

Complementing the high-level protocol outlined in §4.2, this appendix details the specific hyperparameter settings, engineering constraints, and module-level configurations required to reproduce our results.

B.1 Hyperparameters and Optimization

Across all experiments, we utilized the Hugging Face `peft` and `transformers` libraries.

- **Optimization:** We employed `Adafactor` for T5-based architectures (mT5) to manage memory efficiency, and `AdamW` for all other models.
- **Learning Rate Strategy:** We adopted a dynamic learning rate scheduler that decays inversely relative to the shot size (k). For low-resource settings ($k \leq 20$), we utilized a higher initial LR to allow rapid adaptation of the LoRA modules. For higher shot settings ($k \geq 100$), the LR was reduced to preserve the priors learned during the English fine-tuning stage.
- **Dynamic LoRA Rank:** To balance plasticity and stability, the LoRA rank r was not fixed but determined by $r = \min(64, \max(8, 2 \times \log_2(k+1)))$.
- **Random Seeds:** To mitigate data sampling variance in few-shot scenarios, the specific training subsets ($k \in \{0, 5, 20, 100, 500\}$) were extracted using three distinct random seeds (42, 43, 44). This sampling protocol accounts for the over 400 experimental runs aggregated in our statistical analysis.

B.2 Architecture-Specific Engineering Nuances

We implemented specific technical adjustments to address the distinct behaviors of each architecture family:

Language Code Forcing (mBART). Unlike other models, mBART-50 relies heavily on language codes. A critical implementation detail for cross-lingual transfer is the decoupling of encoder and decoder language tags. During the target adaptation phase, we forcefully set the **Encoder** source language to the target language code (e.g., `de_DE`), while locking the **Decoder** target language to English (`en_XX`). This forces the model to align foreign input representations into the English label space.

Instruction Masking (Bloomz, Gemma). For Decoder-only models trained on causal language modeling, we applied Instruction Masking to the loss computation. By setting the labels for the input prompt tokens to -100, we ensure that gradients are backpropagated solely based on the generated sentiment label, preventing the model from memorizing the instruction template. Furthermore, we

employed Left Padding during batch training to ensure correct position embeddings for the generated tokens.

Tokenizer Hacking (mBERT, XLM-R). To mitigate the impact of OOV fragmentation described in the main text:

- For **mBERT**, we manually injected unused tokens (e.g., `[unused1]`) to serve as language-specific markers (e.g., `[MASC]`, `[FEM]` for French attributes) and explicitly added a `[CJK]` marker to reinforce character boundaries in Chinese.
- For **XLM-R**, we used a custom `TemplateProcessing` pipeline to correct the inconsistent insertion of <s> and </s> tokens observed when processing German and French sentence pairs via SentencePiece.

Output Parsing Logic. For generative models (mT5, mBART, Bloomz, Gemma), we set `predict_with_generate=True`. To handle generation artifacts (e.g., hallucinations or repeated prefixes), we implemented a regex-based factory function that extracts the first occurrence of valid label tokens ("Positive", "Negative", "Neutral") and discards the remainder of the sequence.

B.3 LoRA Module Targeting

The precise modules targeted for Low-Rank Adaptation differed by architecture to maximize efficiency:

- **mBERT/XLM-R:** `query`, `key`, `value`, `dense`. Notably, for XLM-R in high-resource settings ($k > 20$), we dynamically expanded targets to include `output.dense`.
- **mT5/mBART:** `q`, `v`, `k`, `o`, `wi`, `wo`, `fc1`, `fc2`.
- **Bloomz:** `query_key_value` (fused), `dense`, `dense_h_to_4h`, `dense_4h_to_h`.
- **Gemma-3:** All linear projections: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`.

B.4 Prompt Configurations

Since the Decoder-only and Encoder-Decoder architectures operate on a text generation paradigm, the specific formatting of input prompts is critical for inducing the correct label space. We employed distinct prompting strategies tailored to the pre-training nature of each model family.

Instruction-Tuned Models (Bloomz, Gemma-3). For these Causal LMs, we utilized a rigid instruction template to constrain the generation to the valid label set. The prompt explicitly defines the semantic meaning of the labels and enforces a strict output format to minimize hallucinations. The exact template is as follows:

```
Classify the sentiment of this product
review as "Positive", "Negative", or
"Neutral".
- Positive: Expresses favorable emotions
(e.g., joy, satisfaction).
- Negative: Expresses unfavorable
emotions (e.g., anger, disappointment).
- Neutral: States a fact without emotion.
Only respond with exactly one of these
words and nothing else: Positive,
Negative, Neutral.
### Input: {review}
### Output:
```

Seq2Seq Models (mT5, mBART). Unlike the instruction-tuned models, mT5 and mBART were pre-trained with shorter, task-specific prefixes (following the T5 paradigm). Consequently, we employed a concise prefix-style prompt rather than a verbose system instruction.

- **Input Format:** `sentiment analysis: {review}`
- **Target Output:** The model is trained to generate the raw label tokens (e.g., `Positive`) directly.

This minimal prompting strategy aligns with the T5 framework's original design for multi-task learning, effectively switching the model's internal state to the sentiment classification task embedding.

B.5 Model Assets

Table 6 lists the specific model checkpoints utilized. We provide the direct Hugging Face URLs to ensure the exact versions are used for reproducibility.

C Statistical Supplement for LMEM Analysis

This appendix provides the comprehensive statistical deliverables supporting the attribution analysis in §5.2. We present the regression tables for the key model specifications, detailed collinearity diagnostics via Variance Inflation Factors (VIF), and validation of the LMEM assumptions.

Model	Params	Hugging Face Checkpoint URL
mBERT	110M	https://huggingface.co/bert-base-multilingual-cased
XLM-R Base	250M	https://huggingface.co/xlm-roberta-base
XLM-R Large	560M	https://huggingface.co/xlm-roberta-large
mT5 Base	580M	https://huggingface.co/google/mt5-base
mBART-50 large	610M	https://huggingface.co/facebook/mbart-large-50
Bloomz	560M	https://huggingface.co/bigscience/bloomz-560m
Gemma-3	270M	https://huggingface.co/google/gemma-3-270m-it

Table 6: Official checkpoints for the SMLMs evaluated in this study.

C.1 Full Regression Results

To ensure reproducibility, we present the detailed regression results for the pivotal model specifications discussed in the main text. Each table details the standardized coefficients (β), p-values, and 95% confidence intervals for predictors added in a sequential, stepwise manner. The final column reports the cumulative AIC, allowing for a direct assessment of how each additional variable improves the model's trade-off between goodness-of-fit and complexity. A progressively decreasing AIC indicates that the newly added variable contributes significant explanatory power.

Table 7 details the results for Model M7, which we identify as the most robust specification. In this model, we control for Resource Scale (X_{shot}), Syntactic Distance (D_{syn}), and Emotion Distance (D_{emo}).

Predictor	Coef. (β)	P-val	[0.025	0.975]	AIC
Intercept β_0	0.657	0.000	0.609	0.704	NULL
X_{shot}	0.008	0.000	0.004	0.012	-364.54
D_{syn}	0.029	0.024	0.004	0.054	-392.44
D_{emo}	-0.046	0.000	-0.064	-0.027	-408.31
D_{cult}	**-0.016**	**0.046**	**-0.032**	**-0.000**	**-410.24**

Table 7: Regression Results for Model M7. Cultural Distance (D_{cult}) remains a significant negative predictor ($p < 0.05$) distinct from linguistic confounders. The AIC column tracks cumulative model fit; its progressive decrease confirms that D_{cult} contributes unique explanatory power, optimizing the overall trade-off between complexity and accuracy. Notably, the positive coefficient for Syntactic Distance likely reflects Transformers' inherent robustness to structural reordering via self-attention mechanisms.

Regression results of all models. The results for combinatorial specifications (e.g., controlling for OOV Rate, Orthographic Distance) are available in the supplementary material repository, namely Table 8 to 13.

Predictor	Coef. (β)	P-val	[0.025	0.975]	AIC
Intercept β_0	0.657	0.000	0.609	0.704	NULL
X_{shot}	0.008	0.000	0.004	0.012	-364.54
D_{cult}	**-0.024**	**0.000**	**-0.033**	**-0.015**	**-388.33**

Table 8: Regression Results for Model M1.

Predictor	Coef. (β)	P-val	[0.025	0.975]	AIC
Intercept β_0	0.657	0.000	0.609	0.704	NULL
X_{shot}	0.008	0.000	0.004	0.012	-364.54
D_{geo}	-0.031	0.001	-0.049	-0.012	-398.34
D_{cult}	**0.003**	**0.746**	**-0.016**	**0.022**	**-396.44**

Table 9: Regression Results for Model M2.

C.2 Collinearity Diagnostics and VIF Analysis

A critical challenge in cross-cultural attribution is the high correlation between Cultural Distance and other typological factors (e.g., geographically distant languages often have distinct scripts or higher OOV Rate). To quantify this multicollinearity, we computed the VIF for key predictors in our LMEM. High VIF scores (typically > 5) indicate that a predictor is highly correlated with other predictors in the model, potentially inflating standard errors and obscuring independent effects.

As shown in Table 14, the presence of high VIF scores for cultural and typological features directly explains the positive correlations observed in some preliminary models (e.g., when controlling for OOV Rate or Geographic Distance alone). These variables are so strongly correlated with performance degradation that they mathematically absorb the variance of Cultural Distance. This phenomenon, known as statistical suppression, highlights the challenge of isolating culture's effect when it is intrinsically entangled with other observable typological features. Our final model selec-

Predictor	Coef. (β)	P-val	[0.025	0.975]	AIC
Intercept β_0	0.657	0.000	0.609	0.704	NULL
X_{shot}	0.008	0.000	0.004	0.012	-364.54
D_{orth}	0.030	0.000	0.019	0.042	-413.60
D_{cult}	**-0.003**	**0.578**	**-0.014**	**0.008**	**-411.91**

Table 10: Regression Results for Model M4.

Predictor	Coef. (β)	P-val	[0.025	0.975]	AIC
Intercept β_0	0.657	0.000	0.609	0.704	NULL
X_{shot}	0.008	0.000	0.004	0.012	-364.54
D_{syn}	0.029	0.024	0.004	0.054	-392.44
D_{emo}	-0.046	0.000	-0.064	-0.027	-408.31
D_{cult}	**-0.016**	**0.046**	**-0.032**	**-0.000**	**-410.24**

Table 11: Regression Results for Model M7.

Predictor	Coef. (β)	P-val	[0.025	0.975]	AIC
Intercept β_0	0.657	0.000	0.609	0.704	NULL
X_{shot}	0.008	0.000	0.004	0.012	-364.54
D_{syn}	-0.019	0.024	-0.036	-0.003	-392.44
D_{cult}	**-0.008**	**0.347**	**-0.025**	**0.009**	**-391.32**

Table 12: Regression Results for Model M14.

Predictor	Coef. (β)	P-val	[0.025	0.975]	AIC
Intercept β_0	0.657	0.000	0.609	0.704	NULL
X_{shot}	0.008	0.000	0.004	0.012	-364.54
R_{oov}	-0.054	0.000	-0.076	-0.032	-404.90
D_{cult}	**0.026**	**0.020**	**0.004**	**0.047**	**-408.20**

Table 13: Regression Results for Model M22.

tion (M7) deliberately omits these high-collinearity variables to reveal the true, independent impact of culture.

This observation reinforces our theoretical framework: Cultural misalignment functions as a latent high-level bottleneck. When surface-level noise (OOV Rate) is overwhelmingly strong, it masks the subtle pragmatic signal; however, when structural factors (Syntax, Emotion) are controlled without introducing collinear surface proxies, the independent negative impact of culture becomes statistically visible and robust.

C.3 LMEM Structure and Random Effects Diagnostics

We employed a LMEM with a random intercept for each model architecture. The mathematical specification is:

$$y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + u_i + \epsilon_{ij} \tag{3}$$

where $u_i \sim \mathcal{N}(0, \sigma_u^2)$ represents the random effect for Model i.

Necessity of Random Effects. We incorporated random intercepts for each model architecture to strictly control for the inherent performance heterogeneity across different SMLMs. A likelihood ratio test confirmed that modeling this hierarchical structure is statistically superior to a fixed-effects-only approach. By accounting for model-specific baselines, we ensure that the analysis isolates the variability driven by cross-lingual distance, rather than conflating it with the general architectural superiority (e.g., the consistent baseline advantage of XLM-R over smaller decoder-only models). This disentanglement prevents the misattribution of general model robustness to cultural adaptability.

C.4 Post-Hoc Power Analysis via Monte Carlo Simulation

A potential methodological concern regarding our hierarchical LMEM analysis is that the limited number of macroscopic target languages ($N = 4$; FR, GE, ZH, JA) might constrain the independent variation needed for reliable statistical attribution. To rigorously validate the statistical robustness of our model against this limited macroscopic sample size, we conducted a post-hoc power analysis using Monte Carlo simulations.

While the number of distinct cultural profiles is small, our experimental framework leverages a highly nested, repeated-measures design across multiple models ($N = 7$), few-shot settings ($N = 5$), and random seeds ($N = 3$). This structural design generates a substantial number of independent observations ($7 \times 4 \times 5 \times 3 = 420$ transfer pairs) that significantly amplify statistical power through partial pooling in the mixed-effects architecture.

To quantify this, we simulated 1,000 synthetic datasets based on the empirical structure of our most robust specification (Model M7). For each iteration, we generated simulated sentiment transfer scores ($F1$) by injecting the empirical fixed-effect estimates (most critically, $\beta_{cult} = -0.016$, alongside specific control parameters), model-specific random intercepts derived from the empirical group variance ($\sigma_u^2 \approx 0.004$), and normally distributed residual noise ($\sigma_\epsilon^2 \approx 0.0024$). We then refit the LMEM to each simulated dataset and computed the proportion of models that successfully detected

Predictor	VIF Score	Collinearity
D_{cult}	**13.30**	**High**
D_{orth}	**8.48**	**High**
R_{oov}	**30.89**	**Very High**
D_{geo}	**9.84**	**High**

Table 14: **VIF Scores.** The high VIF values for Cultural Distance, Orthographic Distance, OOV Rate, and Geographic Distance (all > 8) when included in the same model indicate substantial multicollinearity. This explains why the significance of Cultural Distance varies across specifications: these surface and physical variables partially capture or obscure the independent cultural signal. In our final model M7, we select a parsimonious set of predictors with low VIF to isolate culture's distinct impact.

a statistically significant, negative effect for Cultural Distance ($p < 0.05, \beta < 0$).

The Monte Carlo simulation yielded a statistical power of effectively 100% ($1,000$ significant detections out of $1,000$ valid convergences). This empirically confirms that our nested hierarchical design possesses ample sensitivity to capture the specific impact of cultural misalignment. Consequently, it demonstrates that the observed performance degradation is statistically reliable and not merely an artifact of the small language sample size.

D Supplementary Visualizations and Probing Results

This appendix extends the mechanism analysis from §6.1 by providing visualization evidence across diverse architectures and conducting a rigorous control experiment to validate the robustness of the cultural diagnostic probe.

D.1 Latent Space Visualization across Architectures

In the main text, we visualized the embedding space of mBERT to demonstrate the Pragmatic Alignment Paradox. To verify that this geometric disjointedness is not specific to encoder-only models, we projected the latent representations of the Encoder-Decoder (mT5) and Decoder-only (Gemma-3) architectures using the same t-SNE settings.

As shown in Figure 6, the geometric separation between Western (FR, DE) and Eastern (ZH, JA) clusters remains pervasive. Despite the fundamental differences in attention mechanisms and training objectives, all SMLMs encode these cultural groups into isolated subspaces. This universality supports our claim that the bottleneck is intrinsic to the multilingual representation alignment rather than a model-specific artifact.

D.2 Cross-Lingual Probe Transfer

A potential critique of our diagnostic probe's high accuracy ($> 99\%$ in §6.1) is that the classifier might rely on superficial heuristics, specifically **lexical memorization** (e.g., specific English tokens) or **script discrepancies** rather than capturing a shared latent cultural encoding.

To rule out the hypothesis that the probe merely memorizes language-specific vocabulary, we conducted a **Cross-Lingual Probe Transfer** experiment. We adopted a Leave-One-Pair-Out protocol:

- **Training Phase:** We trained the logistic regression probe solely on representations from EN (Western) and ZH (Eastern). The probe learns to identify a decision boundary separating these two specific languages.
- **Testing Phase (Zero-Shot):** We froze the probe and evaluated its classification accuracy on an unseen language pair: DE, FR (Western) vs. JA (Eastern).

Model Architecture *(Train: EN/ZH)*	In-Domain *Test: EN/ZH*	Zero-Shot *Test: DE+FR / JA*
mBERT (Enc.)	100.0%	**90.16%**
mT5 (Enc-Dec.)	100.0%	**93.70%**
Gemma-3 (Dec.)	100.0%	**99.87%**

Table 15: Cross-Lingual Probe Transfer Accuracy. The probe is trained solely on EN (West) vs. ZH (East). The high zero-shot accuracy on the unseen language set (DE/FR vs. JA) confirms that the Western and Eastern subspaces are geometrically consistent across languages and architectures, ruling out simple lexical memorization.

Rationale. English and German have distinct vocabularies (despite some cognates), as do Chinese and Japanese (excluding shared Kanji/Hanzi, which typically have different encodings). If the probe relies on memorizing English token IDs, it should fail to classify German samples correctly. Conversely, if the probe succeeds on the unseen pair, it proves that the model maps these languages into a shared geometric direction of cultural variance.

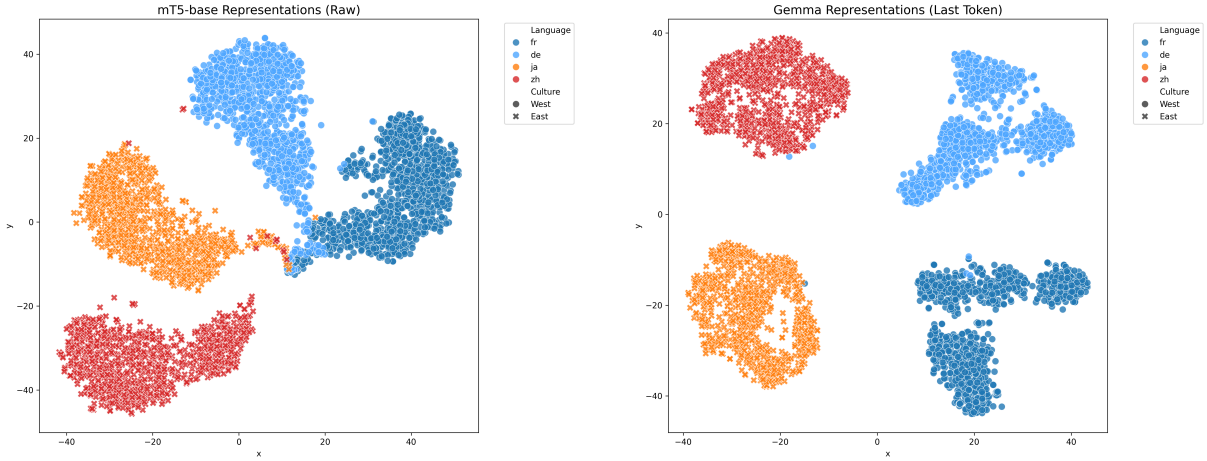

Figure 6: t-SNE Projections for mT5 and Gemma-3. Consistent with the mBERT visualization in the main text, both architectures exhibit a stark geometric separation between the Western and Eastern clusters. This confirms that the cultural disjointedness is an architecture-agnostic phenomenon, persisting regardless of the pre-training objective.

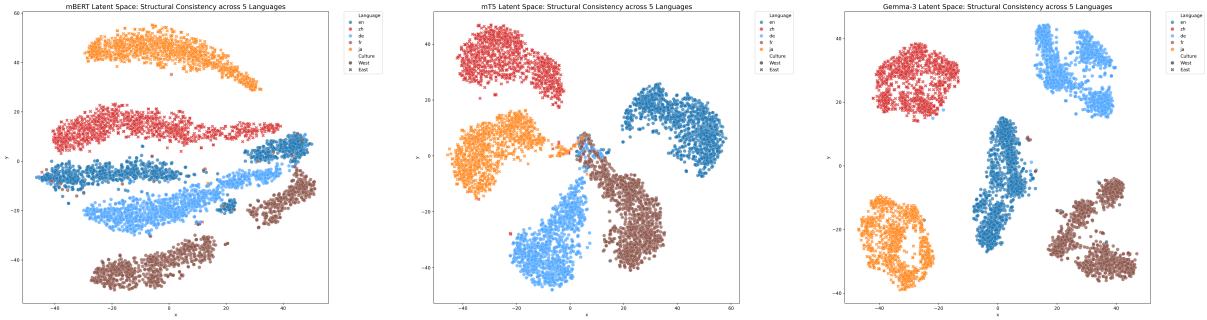

Figure 7: Latent Space Visualization of 5 Languages across Architectures. These t-SNE projections verify the structural consistency hypothesis. In all three architectures, the unseen Western languages (DE, FR) spontaneously cluster with the source Western language (EN), while the unseen Eastern language (JA) aligns with the source Eastern language (ZH). This geometric alignment explains the high zero-shot transfer accuracy of the diagnostic probe.

Results and Discussion. As presented in Table 15, the probe maintains high classification accuracy on the unseen DE, FR/JA pair. This quantitative finding is visually corroborated by Figure 7, where we observe that German and French vectors naturally reside in the same relative Western subspace as English, while Japanese vectors align with the Chinese subspace. This result is critical: it demonstrates that the geometric boundary between Western and Eastern languages is structurally consistent across the multilingual embedding space. The model places German vectors in the same relative Western subspace as English, and Japanese vectors in the same Eastern subspace as Chinese. This confirms that the observed separability is driven by deep representation alignment (or misalignment) rather than superficial token memorization, further validating the Pragmatic Alignment Paradox.

E High-Shot Error Persistence

To verify whether the Pragmatic Alignment Paradox (§6.2) is merely a zero-shot artifact, we examined the error distributions of the 500-shot models for Eastern targets (ZH, JA). Our analysis confirms that the representative pragmatic failures detailed in Table 5 (e.g., Concessive Buffering, Politeness Masking) stubbornly persist even after exposure to 500 in-domain examples.

This persistence indicates that while few-shot adaptation successfully aligns explicit polarity lexicons, it fails to overwrite the strong English Culture Prior established during source fine-tuning (§4.2). The models continue to over-attend to localized surface tokens rather than decoding the implicit illocutionary force. This qualitative finding mechanically explains the Cultural Cliff observed in Figure 2 (§5.1): brute-force data scaling in the few-

shot regime is structurally insufficient to bridge deep-seated cultural misalignment.

F Computation of Emotion Distance

To ensure full reproducibility of the LMEM specifications (particularly Model M7), this section details the computation of the Emotion Distance metric (D_{emo}). As introduced in §3.2, D_{emo} quantifies the semantic divergence of emotional concepts between the source language (L_S, English) and a given target language (L_T).

Following the theoretical framework of Jackson et al. (2019), the emotional lexicon of a language can be represented as a semantic network derived from *colexification*, defined as cases where a single word expresses multiple distinct concepts. For a standardized set of core emotion concepts, we represent each language L as a semantic proximity matrix $\mathbf{M}^{(L)}$, where each element $M^{(L)}_{u,v}$ captures the empirical similarity between emotion concepts u and v within that language's lexicon.

To calculate the pairwise distance between languages, we extract the flattened, upper-triangular vector of these semantic matrices, yielding a high-dimensional emotional profile vector $\mathbf{e}^{(L)}$ for each language. The Emotion Distance is then operationalized as the cosine distance between the vectors of the source and target languages:

$$D_{emo}(L_S, L_T) = 1 - \frac{\mathbf{e}^{(L_S)} \cdot \mathbf{e}^{(L_T)}}{\|\mathbf{e}^{(L_S)}\|_2 \|\mathbf{e}^{(L_T)}\|_2} \quad (4)$$

This formulation yields a bounded scalar value $D_{emo} \in [0, 1]$, where higher scores indicate greater structural divergence in how emotional concepts are lexically grouped and conceptualized. These raw distances were subsequently Z-score normalized before being integrated as predictors in the LMEM (§3.3) to ensure comparability with other variables.

G Per-Language Statistics for Hierarchical Variables

Table 16 presents the exact per-language values for all hierarchical control and target variables utilized in our statistical attribution framework (§3.2). All distance metrics are computed relative to the source language (EN).

Target Language	$\mathbf{D_{cult}}$	$\mathbf{D_{geo}}$	$\mathbf{R_{oov}}$	$\mathbf{D_{orth}}$	$\mathbf{D_{syn}}$	$\mathbf{D_{emo}}$
French (FR)	71.610	0.000	0.283	0.044	0.460	0.619
German (DE)	57.000	0.100	0.289	0.026	0.420	0.703
Chinese (ZH)	100.657	1.000	0.921	0.011	0.570	0.805
Japanese (JA)	91.640	0.500	0.950	0.013	0.660	0.881

Table 16: Per-language statistics for all variables. D_{cult}: Cultural Distance; D_{geo}: Geographic Distance; R_{oov}: OOV Rate; D_{orth}: Orthographic Distance; D_{syn}: Syntactic Distance; D_{emo}: Emotion Distance.