Towards Style Alignment in Cross-Cultural Translation

Shreya Havaldar*, **Adam Stein***, **Eric Wong, & Lyle Ungar**
University of Pennsylvania
{shreyah,steinad,exwong,ungar}@seas.upenn.edu

Abstract

Successful communication depends on the speaker's *intended style* (i.e., what the speaker is trying to convey) aligning with the listener's *interpreted style* (i.e., what the listener perceives). However, cultural differences often lead to misalignment between the two; for example, politeness is often lost in translation. We characterize the ways that LLMs fail to translate style — biasing translations towards neutrality and performing worse in non-Western languages. We mitigate these failures with RASTA (Retrieval-Augmented STylistic Alignment), a method that leverages learned stylistic concepts to encourage LLM translation to appropriately convey cultural communication norms and align style.[1]

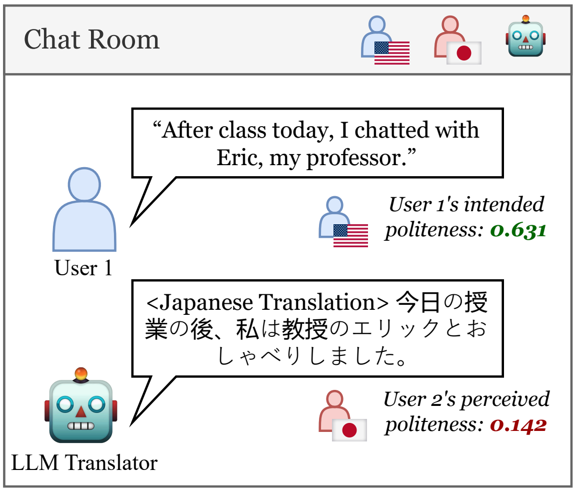

Figure 1: An example of cross-cultural communication facilitated by an LLM. User 1 (American) intends to be polite, but User 2 (Japanese) interprets the message as slightly impolite, given Japanese cultural norms don't typically condone calling a professor by their first name.

1 Introduction

People from different cultural backgrounds must interact as our world grows more interconnected. Machine translation helps promote such intercultural dialogue (Shadiev and Huang, 2016; Khasawneh, 2023), as shown in Figure 1, with LLMs being increasingly adopted to facilitate translation (Albarino, 2024). These models bridge the *linguistic* gap that may arise during communication; however, another important gap to address is the *cultural* gap (Hershcovich et al., 2022).

Domains like healthcare and education benefit greatly from shared knowledge (Lee, 2023); however, communication practices differ across cultures (Schouten and Meeuwesen, 2006; Hofstede, 1986). These differences introduce unique challenges to cross-cultural communication (Moorjani and Field, 1988). For instance, statements that are helpful and appropriate in one culture could be interpreted as critical in another (Hall, 1976).

Successful communication relies on a speaker's intended style to align with a listener's interpreted style[2] (Thomas, 1983; Tannen, 1983). However, cultural differences in individuals' thoughts and actions (Lehman et al., 2004) can lead to a mismatch of such intent and interpretation, so, to successfully bridge the cultural gap, LLMs must translate *style* along with *content* (i.e. the literal meaning).

In this work, we analyze and mitigate translation errors arising from this cultural gap. We focus on style, a key component of communication, and find that modern LLMs often destroy style during translation (Kajava et al., 2020; Troiano et al., 2020).

For instance, in Figure 1, an American (User 1) refers to her professor by his first name. User 1 intends to be polite, and the LLM facilitating the

*Equal contribution

[1]Code, data, & prompts available at `https://github.com/shreyahavaldar/style_alignment`

[2]Linguistic style reflects the systematic variation in linguistic choices across different contexts and speakers, i.e. features of grammar and vocabulary that signal social identity, attitude, and communicative intent (Biber and Conrad, 2009).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32213–32230
July 27 - August 1, 2025

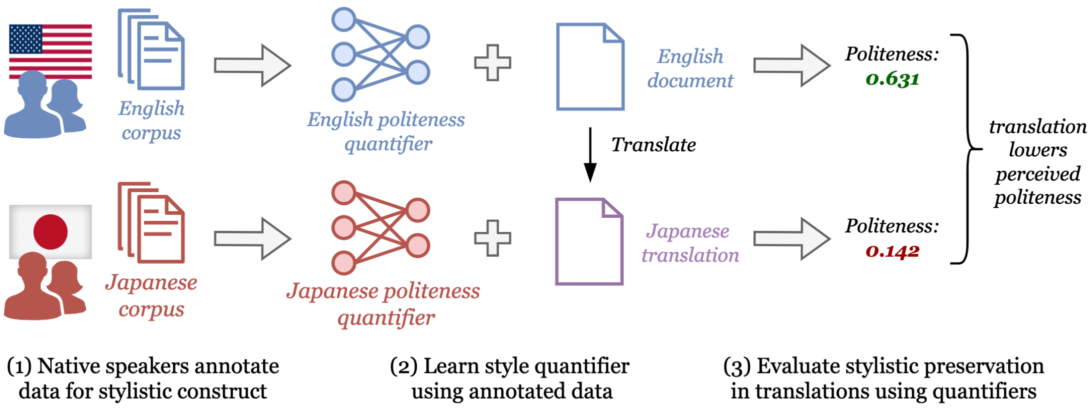

Figure 2: Evaluating how translation affects style. We first select a multilingual style corpus (e.g. the Holistic Politeness Corpus from Havaldar et al. (2023)) annotated by native speakers from corresponding cultures. We then train separate style quantifiers for each language using text annotated by native speakers. Using these style quantifiers to label the style of the translated text, we can measure how well style is preserved during translation.

interaction translates the message's content. However, when the Japanese listener (User 2) reads the message, she interprets it as impolite, given that Japanese cultural norms discourage referring to professors by their first names.

In this work, we characterize failures in style alignment of LLMs. To mitigate these failures, we develop a method where we learn stylistic concepts and leverage available native text to align style during translation without degrading translation quality. Our contributions are as follows:

- We introduce *style alignment* as a goal for cross-cultural translation.
- Using a variety of datasets and LLMs, we characterize the following style alignment failures:
  1. LLMs struggle to translate style, and perform worse in non-Western languages.
  2. LLMs bias translations towards a neutral style, reducing real-world variance.
  3. Current translation metrics poorly reflect style alignment.
- We present RASTA (Retrieval-Augmented STylistic Alignment), a method to *preserve style during translation*, generating translations preferred by bilingual speakers.

2 Designing a Metric for Style Alignment

In a world where translation models perfectly take into account cultural differences in style, a text's *intended* style in Language $\mathcal{L}_1$ should match the translated text's *interpreted* style in language $\mathcal{L}_2$.

Using this assumption, we quantify the degree of style alignment, $\mathcal{A}$, by measuring the difference between the style of a text in language $\mathcal{L}_1$ and the

English Examples	**Label (0-1)**
Politeness	
This has already been debated to death.	0.024
Oh I believe this was already debated!	0.465
Intimacy	
Happy New Year!	0.372
Happy new year!!!! Love u	0.681
Formality	
uh, can i have more details pls?	0.044
Could you provide more details, please?	0.989

Table 1: Examples of text with the same content, but different style labels from our three style datasets.

style of the same text translated to language $\mathcal{L}_2$. For example, we can train two style quantifiers for languages $\mathcal{L}_1$ and $\mathcal{L}_2$, which output a number between 0 and 1, measuring the degree of style expressed in an input text. Formally, our style quantifiers are mappings, $\mathcal{C}_1 : \mathcal{L}_1 \rightarrow [0, 1]$ and $\mathcal{C}_2 : \mathcal{L}_2 \rightarrow [0, 1]$. Then, given a text x, a translator $T : \mathcal{L}_1 \rightarrow \mathcal{L}_2$ should ideally satisfy:

$$\forall x \in \mathcal{L}_1, \mathcal{C}_1(x) = \mathcal{C}_2(T(x)). \tag{1}$$

Since this strict equality is unlikely to hold in practice, we measure the extent to which this holds by calculating a style alignment score $\mathcal{A}$, i.e. the product-moment correlation between the style of the original and translated text. The style alignment score $\mathcal{A}$ of a corpus X translated from $\mathcal{L}_1$ into $\mathcal{L}_2$ is calculated as follows:

$$\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2) = r\left(\mathcal{C}_1\left(X_{\mathcal{L}_1}\right), \mathcal{C}_2\left(T\left(X_{\mathcal{L}_1}\right)\right)\right) \tag{2}$$

This value measures how well the intended style of

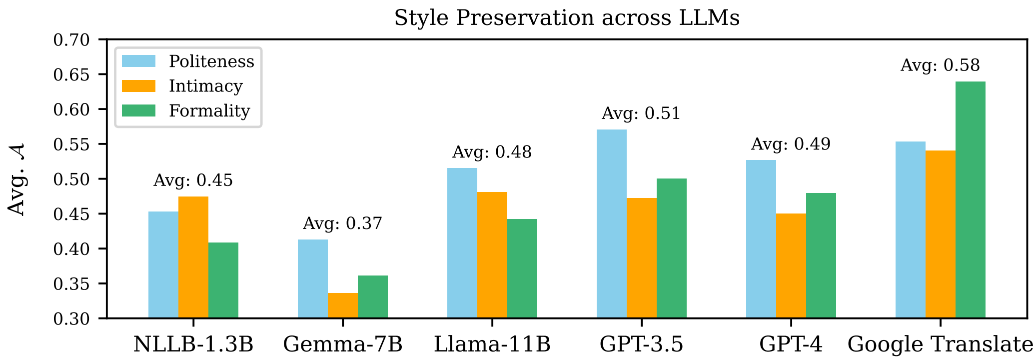

Figure 3: How well do LLMs align style during translation? We show $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ for today's top LLMs, averaged across language pairs. Google Translate is the best at this task, but all models are far from perfect.

a speaker in language $\mathcal{L}_1$ matches the interpretation of a listener in language $\mathcal{L}_2$.

2.1 Selecting Style Datasets

In additional to an alignment metric, we also require data. Specifically, we require style datasets that *span multiple languages*, and are *annotated by native speakers*.

We select three open source datasets meeting these criteria:

1. The Holistic Politeness Dataset (Havaldar et al., 2023), containing Wikipedia editor interactions annotated for **politeness** in English, Spanish, Japanese, and Chinese.
2. Multilingual Tweet Intimacy[3] (Pei et al., 2023), containing Facebook posts annotated for **intimacy** in English, Spanish, Brazilian-Portuguese, Italian, French, and Chinese.
3. GYAFC (Rao and Tetreault, 2018) and XFORMAL (Briakou et al., 2021), containing Yahoo Answers annotated for **formality** in English, Brazilian-Portuguese, French, and Italian.

Given these datasets contain samples from different languages *in the same domain*, there is minimal difference between the distribution of translated text and native speaker text. Additional statistics for these three datasets are provided in Table A3.

2.2 Training Style Quantifiers

Automatically measuring intended and interpreted style requires reliable style quantifier models.

For all of our datasets (politeness, intimacy, and formality), we fine-tune a set of Mistral-7B models (Jiang et al., 2023) to output a style score for each sample. We train a separate model per language to encourage understanding of culture-specific stylistic nuances without cross-lingual interference.

The average test RMSE across languages is 0.157, 0.183, and 0.255 for politeness, intimacy, and formality respectively. See Appendix A and Table A2 for more details on style quantifier training and evaluation.

3 Today's LLMs Struggle to Align Style

With the evaluation framework detailed in Figure 2, we can answer the question: *How well do state-of-the-art LLMs translate intended style?*

Selected models. Zhu et al. (2024) evaluates the translation capabilities of widely-used LLMs. We select the top-performing from this work — NLLB-1.3B (El-Alami et al., 2022), GPT-3.5, GPT-4, and Google Translate. We also include two high-performing lightweight LLMs — Gemma-7B (Gemma-Team, 2024) and Llama-3.2-11B-Vision-Instruct (Dubey et al., 2024).

Evaluation setup. Each style dataset contains annotated samples from various languages within a single domain. We use the LLM to translate the samples in one language into the other languages in the dataset using a simple prompt: `Translate the following text from` $\mathcal{L}_1$`to` $\mathcal{L}_2$`: <Text>`, where we provide the source and target languages at runtime. Sampling parameters for all LLMs are provided in Appendix B.

Next, for each language pair in the dataset, we calculate the style alignment $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$. Finally, we average $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ across all language pairs to quantify how well an LLM aligns the given style during translation. Results of this experiment

[3]Intimacy refers to whether a text is a communication between strangers vs. people who are emotionally close.

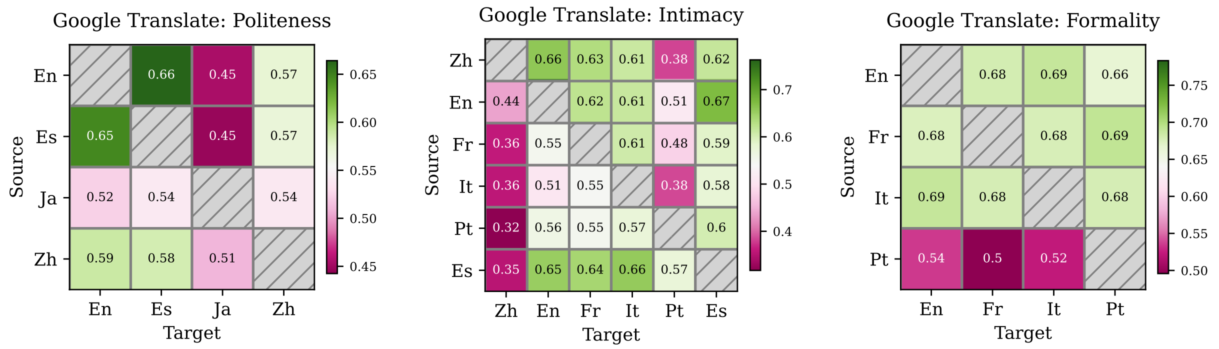

Figure 4: What languages cause Google Translate to struggle the most when aligning style? We show $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ for each language pair; green indicates above average, and pink indicates below average. Results show style alignment is worst in non-Western languages, raising concerns about successful translation in non-Western cultures.

are shown in Figure 3. Google Translate has the highest average $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ of 0.58 across language pairs and styles. GPT-4, GPT-3.5, and Llama-11B all do comparably well, while the smaller models, NLLB-1.3B and Gemma-7B struggle significantly. These results suggest that style alignment can improve with model size, though even the largest models are far from perfect.

3.1 Anglocentric Bias in Style Alignment

Given the suboptimal results in Figure 3, we explore whether the style alignment $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ is higher for certain language pairs than others.

We discuss outputs from Google Translate and GPT-4, as the highest-performing and widest used LLMs. Figure 4 contains heatmaps showing $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ for every source language $\mathcal{L}_1$ and target language $\mathcal{L}_2$). Green indicates a $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ value above average, while pink indicates below average. GPT-4 heatmaps are shown in Figure A1.

We observe a similar pattern across styles: performance is worst when translating into non-Western languages — Japanese (ja), Chinese (zh), and Brazilian-Portuguese (pt). These results suggest a mismatch between intended style and interpretation is most likely when communication involves a non-Western language, highlighting issues for LLM translations used in non-Western cultures.

3.2 LLMs Bias Translations towards Neutrality

Next, we investigate the distribution shift between the style of text written by native speakers of $\mathcal{L}_2$ and the style of text translated to $\mathcal{L}_2$.

We find translations are more "neutral" (i.e. having a label of 0.4-0.6) compared to samples written by native speakers. Additionally, stylistic extremes (i.e. 0-0.1 and 0.9-1) exist in text written by native speakers, but rarely occur in translations.

Figure 5 plots the distribution of politeness text written by native speakers against English text translated into these languages. We see this neutrality bias in action: the standard deviations of native text (red) are [0.23, 0.20, 0.20] for Spanish, Japanese, and Chinese respectively. However, the translations (blue) have significantly decreased standard deviations of [0.17, 0.09, and 0.13].

This phenomenon is particularly problematic for cross-cultural communication involving heightened emotions like enthusiasm or frustration and may lead to confusion or misunderstandings.

	$\mathcal{A}$ vs. G	$\mathcal{A}$ vs. CK	G vs. CK
Google Trans.	-0.154	-0.548	0.674*
GPT-4	0.243	-0.216	0.702*
GPT-3.5	0.030	-0.396*	0.648*
Llama 3.2	0.070	-0.171	0.788*
NLLB	0.030	-0.270*	0.889*
Gemma	-0.369*	-0.181	0.287*

Table 2: Correlations between our style alignment metric $\mathcal{A}$ and traditional translation metrics: GEMBA (G), and CometKiwi (CK). Results shown are the average across all language pairs $\mathcal{L}_1, \mathcal{L}_2$ and models shown in Figure 3. *indicates $p < 0.05$.

3.3 Modern Metrics Exclude Style Alignment

Given the numerous ways that today's LLMs fail to translate intended style, this raises the question – why don't current translation metrics catch this?

In addition to calculating style alignment, we also calculate two popular reference-free translation metrics: GEMBA (Kocmi and Federmann, 2023) and CometKiwi (Rei et al., 2022) on all

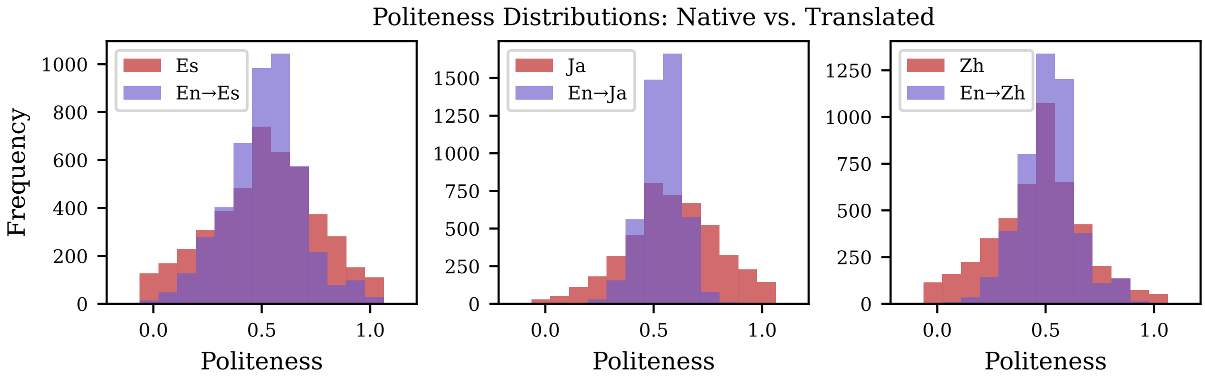

Figure 5: Why do LLMs fail to align style during translation? We plot the politeness distributions of text generated by native speakers vs. translated text. Results show translations skew towards neutral politeness, shrinking standard deviation and reducing the stylistic variance present in the real world.

of our translated data. Finally, we calculate the correlation between these metrics.

Results are shown in Table 2. Across all models, we observe high correlations between the two traditional translation metrics (G and CK), but negative or insignificant correlations when they are compared with $\mathcal{A}$. This indicates traditional metrics poorly evaluate style alignment, providing some insight into why such failures occur.

4 RASTA: Retrieval-Augmented STylistic Alignment

We establish in Section 3 that LLMs fail to properly align style during translation. However, past work has found that LLMs succeed at *classifying* style in a multilingual setting, either via fine-tuning or prompting (Briakou et al., 2021; Plaza-del Arco et al., 2020; El-Alami et al., 2022; Havaldar et al., 2023; Srinivasan and Choi, 2022).

We explore how this stylistic understanding is encoded in embedding space by probing for stylistic concepts and using them to augment the translation process, developing a method that yields more stylistically-aligned results.

4.1 Distinguishing Style in Embedding Space

To begin, we explore how different styles (e.g. polite vs. rude, formal vs. informal) are encoded differently in embedding space.

To determine if different styles within a language are distinguishable in embedding space, we evaluate if samples are clustered by style. To calculate embeddings, we use the BGE-M3 model (Chen et al., 2024a), a recent high-performing multilingual embedding model. Note that in this section we use known style labels to *analyze* the structure of a given embedding space rather than imposing our own structure.

For a corpus $X_{\mathcal{L}}$ representing text in language $\mathcal{L}$ (e.g. Japanese), let $X_{\mathcal{L},\mathcal{S}}$ be the subset of that corpus labeled with style $\mathcal{S}$ (e.g. Japanese text with a politeness label of "slightly rude"). Then, the centroid of texts in language $\mathcal{L}$ with style $\mathcal{S}$ is

$$\mu(\mathcal{L}, \mathcal{S}) = \frac{1}{|X_{\mathcal{L},\mathcal{S}}|} \sum_{x \in X_{\mathcal{L},\mathcal{S}}} \mathcal{E}(x) \qquad (3)$$

where $\mathcal{E}(x)$ is an embedding function. Subfigure (1) within Figure 6 provides a visualization of this.

As seen in Table A1, there is a significant distance between the centroids of *different style* labels within the *same language*, and centroids of the *same style* label in *different languages* (e.g. $\mu(\text{En}, \text{polite}) \rightarrow \mu(\text{En}, \text{rude})$ and $\mu(\text{En}, \text{polite}) \rightarrow \mu(\text{Ja}, \text{polite})$), when compared to the centroids of random subsets of the same size. This distinction implies that LLMs have some encoding of style and how it differs across languages.

To better understand why LLMs succeed at style classification, but fail to translate style, we also calculate the centroids for translated text. Let $\mu(\mathcal{L}_1 \rightarrow \mathcal{L}_2, \mathcal{S})$ denote the centroid of text in $\mathcal{L}_2$ with style $\mathcal{S}$ that has been translated from $\mathcal{L}_1$ to $\mathcal{L}_2$. When we embed this translated text, we observe $\mu(\mathcal{L}_1 \rightarrow \mathcal{L}_2, \mathcal{S})$ is a significant distance away from $\mu(\mathcal{L}_2, \mathcal{S})$. Table A1 suggests a dichotomy between embeddings of text translated into a language $\mathcal{L}_2$ and embeddings of text written by speakers of $\mathcal{L}_2$, despite having the same original style.

Key takeaway. The distinction between translated and native text in embedding space may be a reason for the subpar results we observe in Figure 4. By learning mappings between these concepts, we

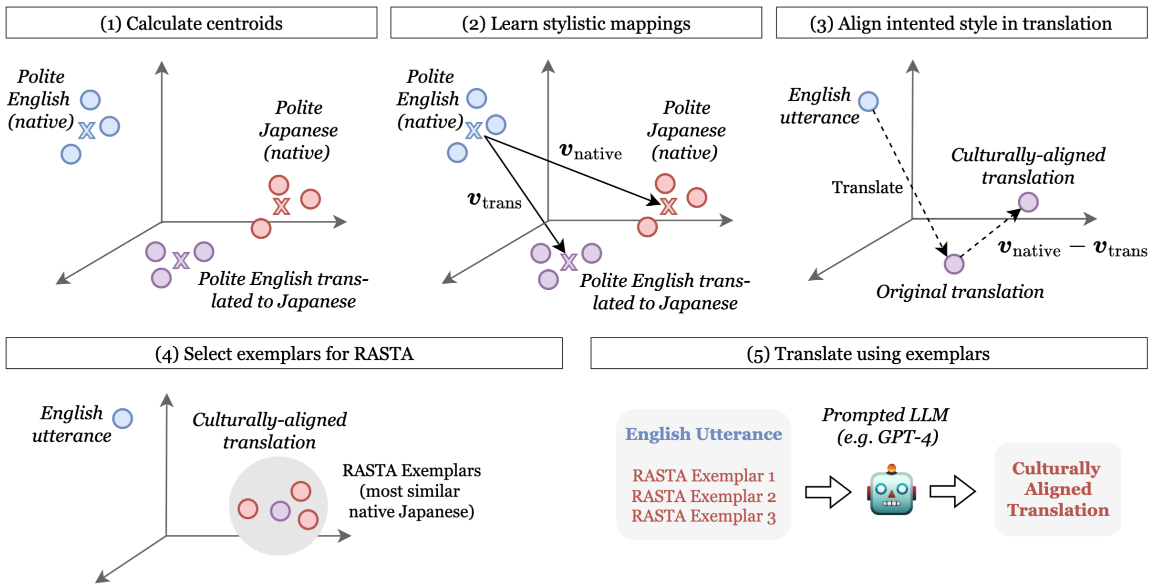

Figure 6: RASTA: our method to align style, shown using English and Japanese. In (1) and (2), we discover stylistic concepts and learn mappings v_{native} from native English $\rightarrow$ native Japanese and v_{trans} from native English $\rightarrow$ translated Japanese. In (3), we apply an alignment mapping $v_{\text{native}} - v_{\text{trans}}$ that aligns an input embedding according to cultural communication norms. In (4) and (5), we use the aligned embedding to select the best few-shot exemplars and generate a more culturally aligned translation. *Note that we perform this process separately for every style and language pair, as centroids and exemplars are unique to style level and language.*

can align the embedding of an input text into one that appropriately reflects the the target style.

4.2 Learning Cultural Alignment Mappings

Formally, for every source language $\mathcal{L}_1$ and target language $\mathcal{L}_2$, we can calculate the mapping from text in $\mathcal{L}_1$ with style $\mathcal{S}$ to text in $\mathcal{L}_2$ with the same style $\mathcal{S}$. This process is detailed below:

$$v_{\text{native}}(\mathcal{L}_1, \mathcal{L}_2, \mathcal{S}) = \mu(\mathcal{L}_2, \mathcal{S}) - \mu(\mathcal{L}_1, \mathcal{S})$$

This gives us a mapping in embedding space that encapsulates *linguistic shift*, e.g. moving from polite text in English to polite text in Japanese.

Similarly, we can calculate the mapping from text in $\mathcal{L}_1$ with style $\mathcal{S}$ to the translations of this text in $\mathcal{L}_2$. This process is detailed below:

$$v_{\text{trans}}(\mathcal{L}_1, \mathcal{L}_2, s) = \mu(\mathcal{L}_1 \rightarrow \mathcal{L}_2, s) - \mu(\mathcal{L}_1, s)$$

This provides an embedding mapping that encapsulates *translation shift*, e.g. moving from polite text in English to its translation in Japanese. Subfigure (2) in Figure 6 shows examples of such mappings.

After calculating v_{native} and v_{trans}, we can calculate exactly how the embedding of a translation in $\mathcal{L}_2$ would need to be transformed to exist in the same part of embedding space as native text in $\mathcal{L}_2$.

In theory, this transformation would eliminate the cultural gap between text in $\mathcal{L}_1$ and its translation in $\mathcal{L}_2$, as Equation 1 would hold. In other words, the intended style in $\mathcal{L}_1$ would exactly match the interpreted style $\mathcal{L}_2$. This cultural alignment mapping, the key component of our RASTA algorithm, can be calculated as follows:

$$v_{\text{align}} := v_{\text{native}}(\mathcal{L}_1, \mathcal{L}_2, \mathcal{S}) - v_{\text{trans}}(\mathcal{L}_1, \mathcal{L}_2, \mathcal{S})$$

Subfigure (3) within Figure 6 details this alignment.

4.3 Preserving Intended Style in Translations

Using this cultural alignment mapping, we can embed and transform any piece of text in source language $\mathcal{L}_1$ to exactly where it *should* lie in embedding space in order for Equation 1 to hold.

Exemplar selection. From there, we find the five exemplars in the training set of $\mathcal{L}_2$ with the highest cosine similarity to this transformed embedding. We use these exemplars as few-shot examples during translation time, encouraging the LLM to analyze how native speakers of $\mathcal{L}_2$ express the style $\mathcal{S}$ of the original text in $\mathcal{L}_1$. The few-shot generation prompt is provided in Appendix D.

Subfigures (4) and (5) within Figure 6 show these final steps of RASTA.

Style	Target	Baseline 1: Vanilla Translation			Baseline 2: "Preserve Style" Prompting			RASTA		
		$\mathcal{A}$↑	CK↑	G↑	$\mathcal{A}$↑	CK↑	G↑	$\mathcal{A}$↑	CK↑	G↑
Politeness	English	0.61	0.80	95.33	0.60	0.80	95.39	0.70	0.78	94.61
	Spanish	0.56	0.75	95.94	0.65	0.75	96.69	0.69	0.75	96.45
	Japanese	0.39	0.80	94.66	0.55	0.80	95.13	0.70	0.78	94.83
	Chinese	0.55	0.76	94.80	0.61	0.76	95.03	0.70	0.75	94.64
	Avg.	0.53	0.78	95.18	0.60	0.78	95.56	0.70	0.77	95.13
	RASTA Δ	+ 32.1%	-1.3%	+0.0%	+16.7%	-1.3%	-0.4%	–	–	–
Intimacy	English	0.64	0.77	93.78	0.62	0.78	94.26	0.66	0.76	93.34
	Spanish	0.62	0.71	94.79	0.58	0.72	95.70	0.59	0.72	95.26
	French	0.38	0.71	93.86	0.57	0.72	94.95	0.60	0.72	94.48
	Italian	0.49	0.72	94.49	0.58	0.72	95.56	0.59	0.72	95.14
	Portuguese	0.29	0.70	95.28	0.45	0.71	95.84	0.46	0.71	95.63
	Chinese	0.28	0.70	92.24	0.37	0.71	93.43	0.39	0.71	93.05
	Avg.	0.45	0.72	94.07	0.53	0.73	94.96	0.55	0.72	94.49
	RASTA Δ	+ 22.2%	+0.0%	+0.0%	+1.7%	+1.4%	-0.5%	–	–	–
Formality	En	0.46	0.82	97.26	0.54	0.82	97.07	0.76	0.81	96.32
	Fr	0.44	0.81	97.11	0.66	0.81	97.35	0.75	0.80	96.98
	It	0.51	0.80	97.73	0.66	0.80	98.01	0.70	0.80	97.83
	Pt	0.50	0.79	97.75	0.72	0.80	97.97	0.78	0.78	97.36
	Avg.	0.48	0.81	97.46	0.64	0.81	97.60	0.75	0.80	97.12
	RASTA Δ	+ 56.3%	-1.3%	-0.4%	+17.2%	-1.3%	-0.5%	–	–	–

Table 3: GPT-4: Evaluation of RASTA with prompting baselines. We measure the style alignment, $\mathcal{A}$, as well as state-of-the-art reference-free translation quality metrics GEMBA (G) (Kocmi and Federmann, 2023) and Comet-Kiwi (CK) (Rei et al., 2022). $\mathcal{A}$ is 1 when the interpreted style (i.e. style of translation) exactly matches the intended style (i.e. style of original text) and 0 when there is no alignment whatsoever. For all metrics, a higher score is better.

5 Experiments & Results

Using our translation framework, RASTA, we re-translate the test sets of all three style datasets. We then evaluate the effectiveness of RASTA by comparing our translations against a set of baselines and measuring both style alignment and traditional translation quality. Since RASTA requires prompting access, we focus on GPT-4 and Llama.

Baseline 1. As described in Section 3, our first baseline uses a *vanilla, straightforward prompt*, providing the LLM with instructions to translate a given text into the target language.

Baseline 2. To evaluate the need for RASTA over sophisticated prompting techniques, we also write a thorough prompt detailing how word-for-word translation may lead to stylistic misalignment, and *explicitly providing instructions to preserve the given style*. The full prompt is in Appendix D.

Evaluation metrics. For RASTA and our two baselines, we calculate $\mathcal{A}$ to measure style alignment and calculate GEMBA and COMETKIWI to measure traditional translation quality.

5.1 RASTA Improves Style Alignment

Table 3 provides results for RASTA using GPT-4. We observe a significant increase in style alignment using RASTA (up to 56% improvement), without significant degradation in GEMBA or COMETKIWI (under 1.5% degradation). We show examples of improved translations in Table A4.

The solid increase in performance from Baseline 2 also indicates that *RASTA better aligns style than what can be done by using a well-written prompt*. We also provide results for Llama-3.2-11B in Table A6, and observe similar patterns of improvement over baselines.

Mitigating Anglocentric bias. As shown in Figure 7, RASTA drastically improves politeness alignment for Chinese and Japanese translation. We observe similar improvement in non-Western languages for intimacy and formality as well (see Figures A1 and A2), highlighting RASTA's ability to de-bias translation performance.

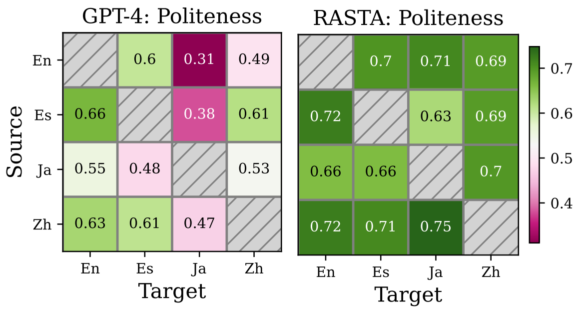

Figure 7: Comparing style alignment between vanilla GPT-4 and RASTA using GPT-4. RASTA improves performance when translating into non-Western languages and reduces the performance gap from 0.35 to 0.12.

Preserving native speaker variance. RASTA also decreases the tendency of translations to be neutral. For GPT-4, the standard deviation of the translated politeness distribution (see Figure 5) increases from vanilla prompting to RASTA as follows: [0.14, 0.10, 0.10] → [0.18, 0.13, 0.15] for Spanish, Japanese, Chinese, a 36% average increase and a much better reflection of the politeness variance occurring in native text.

5.2 Humans Prefer RASTA Translations

To confirm whether native speakers view RASTA translations as an improvement, we run an annotation study on Prolific.

Study setup. For every language pair {English, $\mathcal{L}$} in our politeness and formality datasets[4], we recruit a set of 3 bilingual annotators on Prolific, for a total of 18 annotators across languages and datasets. Annotators are required to have selected $\mathcal{L}$ as their first language and be fully fluent in English. Annotators are then shown 30 randomly selected samples, along with their RASTA translations and their "preserve style" prompting translations, and asked to select which translation better preserves style.

Annotation results. Annotators select RASTA a majority of the time – 61% of the time for politeness, and 63% of the time for formality, indicating RASTA aligns with native speaker intuition. See Appendix F for study details, annotator agreement, and additional results.

Overall, *RASTA succeeds in improving style alignment while still outputting translations that fully preserve content and meaning.*

[4]We do not include intimacy, as a large portion of the dataset contains sexually explicit content.

6 Related Work

Controlling style in generations: Past style transfer work controls for style via sampling and ranking (Niu et al., 2017) and fine-tuning (Rippeth et al., 2022; Garcia et al., 2021). Similar approaches have been used to control style in translation (Niu et al., 2018; Schioppa et al., 2021; Sennrich et al., 2016; Nadejde et al., 2022). RASTA differs from these works as style is not forced on the output; rather, it matches that of the input. TextSETTR (Riley et al., 2021) employs a similar problem setup but only focuses on English, thus eliminating the need for culture-specific style understanding. Low-resource benchmarks (Krishna et al., 2022; Mukherjee et al., 2024) employ language-specific techniques, but therefore lack the generalizability of RASTA.

Culturally-aware translation: Prior work in this space involves benchmark datasets that account for cultural differences in names/objects (Yao et al., 2024; Peskov et al., 2021), geographic locations (Riley et al., 2023), social norms (Huang and Yang, 2023), dialogue (Li et al., 2024a), along with investigations of translating time (Shwartz, 2022), recipes (Cao et al., 2024), and idioms (Li et al., 2024b). However, the techniques described in these works focus on *entity replacement* via fine-tuning or post-processing, thus modifying content. Conversely, RASTA focuses on modifying *style* to increase cultural awareness, keeping content intact.

Retrieval-augmented translation: Recent work uses in-context learning and retrieval-augmented generation (RAG) for LLM translation (Bulte and Tezcan, 2019; Xu et al., 2020; Agrawal et al., 2023; Garcia et al., 2023). Beyond translation pairs, retrieval from external knowledge bases (Conia et al., 2024; Chen et al., 2024b), unstructured data (Wang et al., 2024), and search engines (Gu et al., 2018) have been used to improve translation. RAG has also been combined with multi-step prompting (He et al., 2024). Unlike these approaches, *we do not require translation pairs,* instead used a learned alignment mapping to directly retrieve examples in the target language.

7 Conclusion

We introduce style alignment as a goal for translation, given that success in communication requires an alignment of intended and interpreted style. We characterize the failures of today's LLMs in aligning style: poor performance, Anglocentric bias, skewing towards neutrality, and low correlation with standard translation quality. To mitigate these failures, we introduce RASTA, a method that leverages stylistic concepts and in tandem with native text to generate culturally-aligned translations.

Our work provides essential insights and methodologies to enhance LLMs' capabilities in cross-cultural communication, establishing a foundation for future research in this field.

8 Limitations

In order to calculate our "ground truth" style labels for our alignment metric $\mathcal{A}$, we train models (regressors), which are imperfect. The average RMSE of our style quantifiers (across language pairs and styles) is 0.195. Though this suggests high overall performance, certain predictions may be incorrect. This work would benefit from another annotation study including native speaker annotations of our style quantifier models. Unfortunately, due to resource constraints, multiple annotation studies were not feasible for this work.

We only look at three styles in this work: politeness, intimacy, and formality. Future work is needed to assess whether RASTA improves style alignment for the many other styles that exist. Additionally, content and style are deeply intertwined; it may not always be possible to transform style without also modifying content, making perfect style alignment unachievable.

As with any retrieval-augmented generation method, bias and error in our embedding model may propagate up and result in suboptimal few-shot exemplars. Additionally, since we use cosine similarity to select exemplars, we choose examples close in both content and style, as the embedding encodes both of these constructs. This may provide a suboptimal level of variance in our exemplars.

Though the authors spend time prompt-tuning, we only run experiments with a single prompt for RASTA and each of our baselines; final translations are prompt-dependent and may vary with different prompts. Future work is needed to establish the effect of prompt wording and the number of few-shot exemplars on the final RASTA translations.

This work would be strengthened by explicit inclusion of social and cultural norms to modify translations. However, there are many open questions when determining how to use them for translation: how to distinguish their influence on style vs. content, how to extract norms from style corpora, how to determine which norms to provide to the model, etc. This inclusion is out of scope for this paper and we leave it for future work.

On a higher level, style is subjective, even within languages and cultures, so the ground truth style label likely differs from person to person. This work treats the style of a given text as a static value, which abstracts away all real-world subjectivity.

9 Ethical Considerations

Our definition of "culturally-aware" translation hinges on style alignment; however, culture is deeply complex and consists of many more communication patterns/norms beyond style. While aligning style is a step in the right direction, we acknowledge that it is an incomplete step to full cultural alignment of LLM generations.

We also only study high-resource languages in this work, as we are limited to what languages are available in open-source style datasets. Future work is needed to determine the effectiveness of RASTA on low-resource languages.

Additionally, given we use an LLM to generate the final translations, inherent bias in or fairness concerns associated with the LLM may propagate up into our generated RASTA translations.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Seyma Albarino. 2024. Translation companies accelerating adoption of large language models, alc survey.

Douglas Biber and Susan Conrad. 2009. *Register, genre, and style*. Cambridge University Press.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *57th Annual Meeting of the Association-for-Computational-Linguistics (ACL)*, pages 1800–1809.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Meiqi Chen, Fandong Meng, Yingxue Zhang, Yan Zhang, and Jie Zhou. 2024b. Crat: A multi-agent framework for causality-enhanced reflective and retrieval-augmented translation with large language models. *arXiv preprint arXiv:2410.21067*.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

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

Fatima-zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6048–6056.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning*, pages 10867–10878. PMLR.

Xavier Garcia, Noah Constant, Mandy Guo, and Orhan Firat. 2021. Towards universality in multilingual text rewriting. *arXiv preprint arXiv:2107.14749*.

Gemma-Team. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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

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

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Geert Hofstede. 1986. Cultural differences in teaching and learning. *International Journal of intercultural relations*, 10(3):301–320.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kaisla Kajava, Emily Öhman, Piao Hui, and Jörg Tiedemann. 2020. Emotion preservation in translation: Evaluating datasets for annotation projection. In *Digital Humanities in the Nordic Countries*, pages 38–50. CEUR.

Mohamad Ahmad Saleem Khasawneh. 2023. The potential of ai in facilitating cross-cultural communication through translation. *Journal of Namibian Studies: History Politics Culture*, 37:107–130.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. *Preprint*, arXiv:2110.07385.

Sangmin-Michelle Lee. 2023. The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. *Computer Assisted Language Learning*, 36(1-2):103–125.

Darrin R Lehman, Chi-yue Chiu, and Mark Schaller. 2004. Psychology and culture. *Annu. Rev. Psychol.*, 55:689–714.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024a. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024b. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Angela Moorjani and Thomas T Field. 1988. Semiotic and sociolinguistic paths to understanding culture. In *Toward a new integration of language and culture. Reports of the Northeast Conference on the Teaching of Foreign Languages*, pages 25–45.

Sourabrata Mukherjee, Atul Kr. Ojha, Akanksha Bansal, Deepak Alok, John P. McCrae, and Ondřej Dušek. 2024. Multilingual text style transfer: Datasets & models for indian languages. *Preprint*, arXiv:2405.20805.

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

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. *arXiv preprint arXiv:1806.04357*.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozen Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2023. Semeval-2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2235–2246.

Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. 2021. Adapting entities across languages and cultures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3725–3750, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Carlo Strapparava, L Alfonso Urena Lopez, and M Teresa Martín-Valdivia. 2020. Emoevent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022. Cometkiwi: Istunbabel 2022 submission for the quality estimation shared task. *arXiv preprint arXiv:2209.06243*.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David C Uthus, and Zarana Parekh. 2021. Textsettr: Few-shot text style extraction and tunable targeted restyling. 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 3786–3800.

Parker Riley, Timothy Dozat, Jan A Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. Frmt: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

Elijah Rippeth, Sweta Agrawal, and Marine Carpuat. 2022. Controlling translation formality using pre-trained multilingual language models. *arXiv preprint arXiv:2205.06644*.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara C Schouten and Ludwien Meeuwesen. 2006. Cultural differences in medical communication: a review of the literature. *Patient education and counseling*, 64(1-3):21–34.

Rico Sennrich, Barry Haddow, and Alexandra Birch-Mayne. 2016. Controlling politeness in neural machine translation via side constraints. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40. Association for Computational Linguistics.

Rustam Shadiev and Yueh-Min Huang. 2016. Facilitating cross-cultural understanding with learning activities supported by speech-to-text recognition and computer-aided translation. *Computers & Education*, 98:130–141.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Anirudh Srinivasan and Eunsol Choi. 2022. Tydip: A dataset for politeness classification in nine typologically diverse languages. *arXiv preprint arXiv:2211.16496*.

Deborah Tannen. 1983. *Cross-Cultural Communication.* ERIC.

Jenny A. Thomas. 1983. Cross-cultural pragmatic failure. *Applied Linguistics*, 4:91–112.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354.

Jiaan Wang, Fandong Meng, Yingxue Zhang, and Jie Zhou. 2024. Retrieval-augmented machine translation with unstructured knowledge. *arXiv preprint arXiv:2412.04342*.

Jitao Xu, Josep-Maria Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579. Association for Computational Linguistics.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. *Preprint*, arXiv:2305.14328.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Style Quantifiers: Training Details

To train the style quantifiers, we finetune Mistral-7B (Jiang et al., 2023) on the training datasets for each language in each available style. Since the style labels are normalized values between zero and one, these are regression models rather than classifiers.

Havaldar et al. (2023) find that multilingual vs. monolingual LLMs focus on different features when learning to classify style. Taking this into account, we fine-tune a unique model for each language, as we do not want cross-lingual interference.

Since we train a separate model for each language and style, we train a total of 14 models. Due to resource constraints, we could not finetune the entire Mistral 7B model, so we use QLoRA (Dettmers et al., 2024) to finetune the model on two NVIDIA-A100 GPUs. Specifically, we use 4 bit quantization and a LoRA dimension of 4. We used MSE loss, a learning rate of 5e-5, a batch size of 8, and trained for 5 epochs. The test RMSE of our models is reported in Table A2. We additionally experimented with finetuning the XLM-RoBERTa-Base model (Conneau et al., 2019), but we found that the Mistral-7B model resulted in lower RMSEs for almost all styles and languages.

B LLM Sampling Parameters

For generation from Llama-3.2 and Gemma, we used a temperature of 0.6 and top-p of 0.9. For generation from GPT-3.5 and GPT-4, we used a temperature of 1.0 and top-p of 1.0. We keep all default sampling parameters to evaluate LLMs in their commonly used form.

C Additional Results

Results for our method on the Llama-3.2 11B Vision Instruct model are provided in Table A6.

Discussion of results. Unlike GPT-4, we observe a non-insignificant degradation of translation quality as measured by GEMBA and COMETKIWI. Upon further inspection, this is because Llama-3.2-11B is overly reliant on the few-shot examples; we see many cases where the LLM includes in its translation information present in multiple few-shot examples, but not present in the original text. We use the prompt designed for GPT-4, as we want to report consistent results across LLMs. However, additional prompt tuning and a Llama-specific prompt is likely needed to prevent degradation of translation quality.

Computation	Distance
Different styles within the same language	
$\lvert\mu(\mathcal{L}, \text{polite}) - \mu(\mathcal{L}, \text{rude})\rvert$	2.01±0.26
Identical style across different languages	
$\lvert\mu(\mathcal{L}_1, \text{polite}) - \mu(\mathcal{L}_2, \text{polite})\rvert$	2.91±0.34
$\lvert\mu(\mathcal{L}_1, \text{rude}) - \mu(\mathcal{L}_2, \text{rude})\rvert$	2.77±0.35
Translated vs. native speaker generated within the same style and language	
$\lvert\mu(\mathcal{L}_1 \rightarrow \mathcal{L}_2, \text{polite}) - \mu(\mathcal{L}_2, \text{polite})\rvert$	2.66±0.25
$\lvert\mu(\mathcal{L}_1 \rightarrow \mathcal{L}_2, \text{rude}) - \mu(\mathcal{L}_2, \text{rude})\rvert$	2.32±0.20
Baseline: random subsets of identical size	
$\lvert X_{\text{rand}} - X_{\text{rand}}\rvert$	0.52±0.02

Table A1: Distances in embedding space of politeness dataset subsets. Polite corresponds to the top 20% of politeness scores and rude is the bottom 20% of scores.

Style	Language	Test RMSE
Politeness *(0-1 scale)*	English	0.155
	Spanish	0.164
	Japanese	0.160
	Chinese	0.147
Intimacy *(0-1 scale)*	English	0.127
	Spanish	0.156
	French	0.179
	Italian	0.226
	Portuguese	0.244
	Chinese	0.156
Formality *(0,1 binary)*	English	0.262
	French	0.202
	Italian	0.289
	Portuguese	0.268

Table A2: Test RMSE for our style quantifiers, showing significantly better-than-random performance.

D Prompts

The prompts for baselines and our method is included in Figure A5, Figure A6, and Figure A7.

Prompt development. To come up with the above prompts, we took the following steps:

1. RASTA prompt: We had multiple rounds of iteration, where authors analyzed the results and tried to find qualitative errors (e.g. honorifics incorrectly added, no cultural knowledge incorporated, overcorrection of style, etc.) and modify the prompt accordingly. We settled on a prompt that provided the goal – optimize for style alignment – while still instructing the model to preserve the content.

Dataset	Train Size	Test Size	Contained Languages
Politeness (Havaldar et al., 2023)	18238	2280	English, Spanish, Japanese, Chinese
Intimacy (Pei et al., 2023)	9491	2370	English, Spanish, Portuguese, Italian, French, Chinese
Formality (Rao and Tetreault, 2018; Briakou et al., 2021)	6392	1598	English, French, Italian, Portuguese

Table A3: Dataset statistics. For the three datasets used in this work, we show the number of train and test samples and the included languages.

WARNING: THE EXAMPLES IN THIS TABLE CONTAIN EXPLICIT AND POTENTIALLY OFFENSIVE CONTENT.

Style	Method	Lang.	Sentence	Label (0-1)
Politeness	Source	Ja	なるほど、ここまでの経緯とは別の態度のご発言ということなのでしょうか？ コメント依頼を出したのは私ではないから、外すことを求められても困ります。 図の問題は、私のこだわりではありません。	0.25
	Vanilla	En	I see, so you're saying this comment shows a different attitude from the course of events so far? I'm not the one who requested the comment, so I'm in trouble if I'm asked to remove it. The problem with the diagram is not my concern.	0.566
	RASTA	En	So, are you saying that this statement reflects a different attitude than what's happened up to now? It is rather inconvenient since I didn't ask for the comment to be made, and it's not up to me to decide to eliminate it. The issue with the diagram, it's important to note, is not my personal fixation.	0.277
Intimacy	Source	Es	Con el pitó bien duro #heteros http	0.84
	Vanilla	En	With the whistle blowing hard #heteros http	0.40
	RASTA	En	With a rock hard cock #straight http	0.867
Formality	Source	Pt	Não tenha a ilusão de que ele vai se separar pra ficar com você (são raríssimos esses casos).	0.0
	Vanilla	En	Don't be under the illusion that he is going to leave to be with you (such cases are extremely rare).	1.0
	RASTA	En	Don't kid yourself that he's gonna leave her for you (those cases are super rare).	0.11

Table A4: Qualitative examples comparing RASTA to the vanilla translation baseline with GPT-4. For all examples, RASTA results in a higher level of style alignment than the vanilla translation.

2. Baseline 1: Vanilla Translation Prompt: We used the most straightforward wording that an average user would be most likely to use when asking an LLM to translate text.

3. Baseline 2: "Preserve Style" Prompt: We took our finalized RASTA prompt and removed the few-shot exemplars and style label components, keeping the overview and the instructions fully intact.

E RASTA Examples

Table A4 contains examples comparing RASTA translations to Baseline 1 translations. We include examples which show a large difference between the style of the translation from RASTA and the vanilla baseline.

For politeness, we show a Japanese sentence which is normally translated by GPT-4 to have medium politeness while Japanese native speakers say the source text was rude. Using RASTA, we can see that the translation is now significanly less polite, but still contains the same content.

The intimacy example shows a case where the source sentence contains Spanish slang for male genitalia, which native speakers label as highly intimate. GPT-4 mistranslates this slang to its literal translation, "whistle", with the vanilla method. This is highlighted by the translation's low intimacy label, but using RASTA fixes this issue since the model outputs a translation with the correct intimacy, resulting in the correct slang translation.

The formality example highlights a case where a normal translation is much too formal. Native

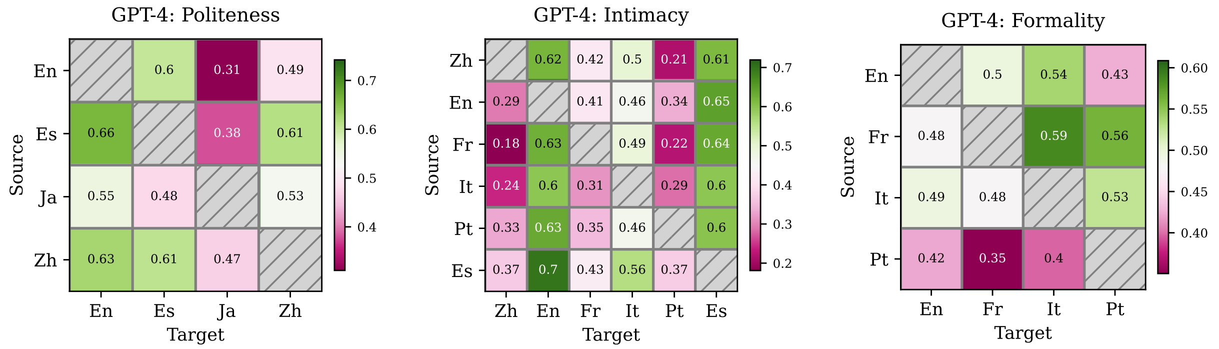

Figure A1: Heatmaps for GPT-4. We show $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ for each language pair; green indicates above average, and pink indicates below average. Results show style alignment is worst in non-Western languages, raising concerns about successful translation in non-Western cultures.

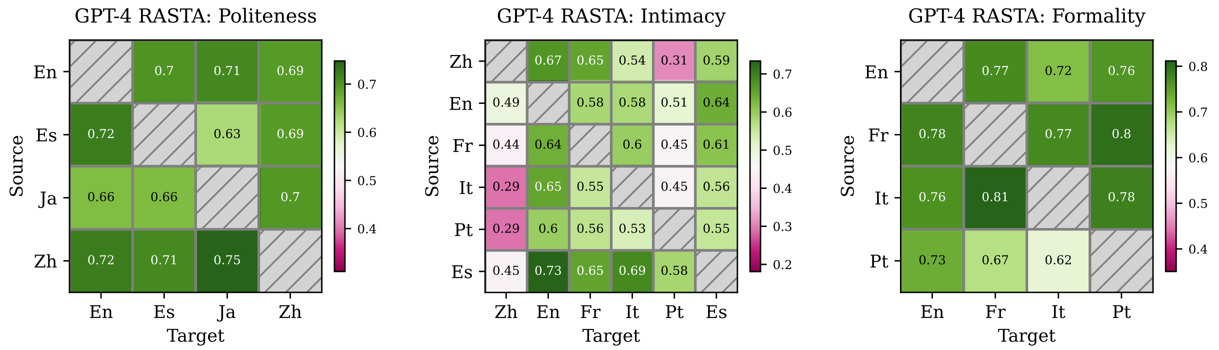

Figure A2: Heatmaps for RASTA with GPT-4. We show $\mathcal{A}(\mathcal{L}_1, \mathcal{L}_2)$ for each language pair; green indicates above average, and pink indicates below average. Results show style alignment is worst in non-Western languages, raising concerns about successful translation in non-Western cultures.

Portuguese speakers identify the source sentence as informal, and RASTA mitigates the erroneous translation by changing the phrase "don't be under the illusion" to "don't kid yourself," an English phrase commonly used in informal settings.

F Human Validation Study

Annotators were sourced bilingual speakers on Prolific and paid $20 an hour. The study took, on average, 18 minutes to complete. Participants annotating for Chinese and Japanese spent 5-6 minutes longer on average than than those annotating for romance languages.

We show one example question asked in our human validation study in Figure A4. We also show the instructions given to the annotators at the start of the survey in Figure A3.

Annotator agreement, as calculated by average pairwise product-moment correlations between annotators is 0.167 for politeness and 0.247 for formality. Agreement is expectedly low due to the highly subjective nature of the task and the low number of samples.

Style	Language	% RASTA Favored
Politeness	Spanish	0.60
	Chinese	0.63
	Japanese	0.60
Formality	Portuguese	0.67
	French	0.63
	Italian	0.60

Table A5: Language-specific results from human validation study. We have three annotators select preferred translation between RASTA and the "preserve style" prompting baseline. We then calculate the majority label for each of the 30 samples per language.

```
Background
Politeness is influenced by social and cultural norms. Because of this, translating
    from one language to another can change politeness levels. A direct word-for-
    word translation does not guarantee that the politeness level of the translated
    text will match the politeness level of the original text.

Instructions
In this survey, you will be asked to analyze translations between English and
    another language and determine which translation better preserves politeness.

Important Note
Keep in mind some translations you see will contain new or modified information to
    reflect the politeness of the original text better.

---------------------

Please select all of the following that are TRUE about politeness. You must get this
     question correct to advance to the rest of the survey!

[] Politeness is a complex construct
[] Politeness is easy to translate
[] Politeness varies across cultures and languages
[] A direct word-for-word translation always preserves politeness
[] A direct word-for-word translation might change politeness
[] A translation that preserves politeness may contain new or modified information
```

Figure A3: Instructions and attention check given to annotators at the start of the survey. Annotators had to pass the attention check (select options 1, 3, 5, and 6) in order to advance to the remainder of the survey.

```
Consider the following translations from English to Spanish. Please select the
    translation that best matches the politeness level of the original text

Original text:
I just came here for the comments. Surprisingly, there are none about hip-hop. The
    ridiculous amount of racist comments and articles debating whether or not hip-
    hop producers are musicians is beyond quantitative measurement.

Translation options:

Option A:
Simplemente vine aquí por los comentarios. Sorprendentemente, no hay ninguno sobre
    hip-hop. La cantidad ridícula de comentarios y artículos racistas debatiendo si
    los productores de hip-hop son o no músicos está más allá de cualquier medición
    cuantitativa.

Option B:
Solo vine aquí por los comentarios. Sorprendentemente, no hay ninguno sobre el hip-
    hop. La cantidad ridícula de comentarios racistas y artículos debatiendo si los
    productores de hip-hop son o no músicos, es más allá de cualquier medición
    cuantitativa.
```

Figure A4: Example of survey question from our human evaluation. Option A in this example corresponds to RASTA, but we randomly shuffle the example order so that there is no pattern to RASTA being Option A or B.

```
Translate the following text from <Source> to <Target>.
Text: <Sample>
Output only the translation.
```

Figure A5: Vanilla translation prompt used for LLMs.

```
Your task is to translate a given piece of text from <Source> to <Target>.
When translating, you must ensure the <Style> level of the translation matches the <
    Style> level of the original text.
Keep in mind that <Style> varies across cultures, so a direct word-for-word
    translation may not always ensure the <Style> level will match that of the
    original text.

This is the text you need to translate:
<Sample>

Now, translate the text so that the translation also has the same <Style> level.
    Output only the translation.
```

Figure A6: "Preserve style" translation prompt used for LLMs.

```
Your task is to translate a given piece of text from <Source> to <Target>.
When translating, you must ensure the <Style> level of the translation matches the <
    Style> level of the original text.
Keep in mind that <Style> varies across cultures, so a direct word-for-word
    translation may not always ensure the <Style> level will match that of the
    original text.

This is the text you need to translate:
<Sample>

This text has a <Style> level of {} out of 1 in {}.

To help you translate the text in a way that preserves <Style>, here are some
    examples of text that have the same <Style> level in {}.
Pay attention to the way <Style> is expressed in these examples, and try to reflect
    it similarly in your translation.
<example 1>

<example 2>

<example 3>

<example 4>

<example 5>

Now, translate the above text to preserve the content of the message, while also
    making sure the <Style> level is similar to the above examples.
Specifically, the translation should also have a <Style> level of {} out of 1 in {}.
     Output only the translation.
```

Figure A7: RASTA translation prompt used for LLMs.

Style	Target	Baseline 1: Vanilla Translation			Baseline 2: "Preserve Style" Prompting			RASTA		
		$\mathcal{A}\uparrow$	CK↑	G↑	$\mathcal{A}\uparrow$	CK↑	G↑	$\mathcal{A}\uparrow$	CK↑	G↑
Politeness	En	0.51	0.77	90.00	0.53	0.76	88.60	0.63	0.72	82.87
	Es	0.61	0.72	91.61	0.61	0.72	91.91	0.64	0.69	89.20
	Ja	0.38	0.74	88.20	0.43	0.74	88.88	0.55	0.70	82.92
	Zh	0.56	0.73	90.52	0.55	0.72	90.49	0.70	0.67	85.32
	Avg.	0.52	0.74	90.08	0.53	0.74	89.97	0.63	0.70	85.08
	RASTA Δ	+ 21.2%	-8.2%	-5.6%	+18.9%	-8.2%	-5.4%	–	–	–
Intimacy	Zh	0.34	0.66	84.02	0.36	0.62	80.02	0.39	0.52	59.10
	En	0.50	0.71	85.34	0.46	0.67	79.92	0.51	0.59	63.70
	Fr	0.55	0.68	85.14	0.51	0.64	80.58	0.60	0.57	64.97
	It	0.55	0.68	86.87	0.49	0.64	81.75	0.57	0.57	63.37
	Pt	0.42	0.67	88.19	0.42	0.64	84.41	0.50	0.55	61.98
	Es	0.53	0.68	87.29	0.46	0.64	82.00	0.56	0.57	63.21
	Avg.	0.48	0.68	86.14	0.45	0.64	81.45	0.52	0.56	62.72
	RASTA Δ	+ 8.3%	-17.6%	-27.2%	+15.6%	-12.5%	-23.0%	–	–	–
Formality	En	0.43	0.80	92.77	0.36	0.79	90.14	0.70	0.68	70.01
	Fr	0.42	0.78	91.74	0.33	0.77	90.09	0.60	0.68	74.85
	It	0.47	0.78	93.14	0.41	0.77	90.92	0.65	0.70	74.87
	Pt	0.46	0.77	92.94	0.42	0.77	91.59	0.62	0.67	72.84
	Avg.	0.44	0.79	92.65	0.38	0.77	90.69	0.64	0.68	73.14
	RASTA Δ	+ 45.5%	-13.9%	-21.1%	+68.4%	-11.7%	-19.4%	–	–	–

Table A6: Llama-3.2-11B: Evaluation of RASTA with prompting baselines. We measure the style alignment, $\mathcal{A}$, as well as state-of-the-art reference-free translation quality metrics GEMBA (G) (Kocmi and Federmann, 2023) and COMETKIWI (CK) (Rei et al., 2022). $\mathcal{A}$ is 1 when the interpreted style (i.e. style of translation) exactly matches the intended style (i.e. style of original text) and 0 when there is no alignment. For all metrics, higher is better.