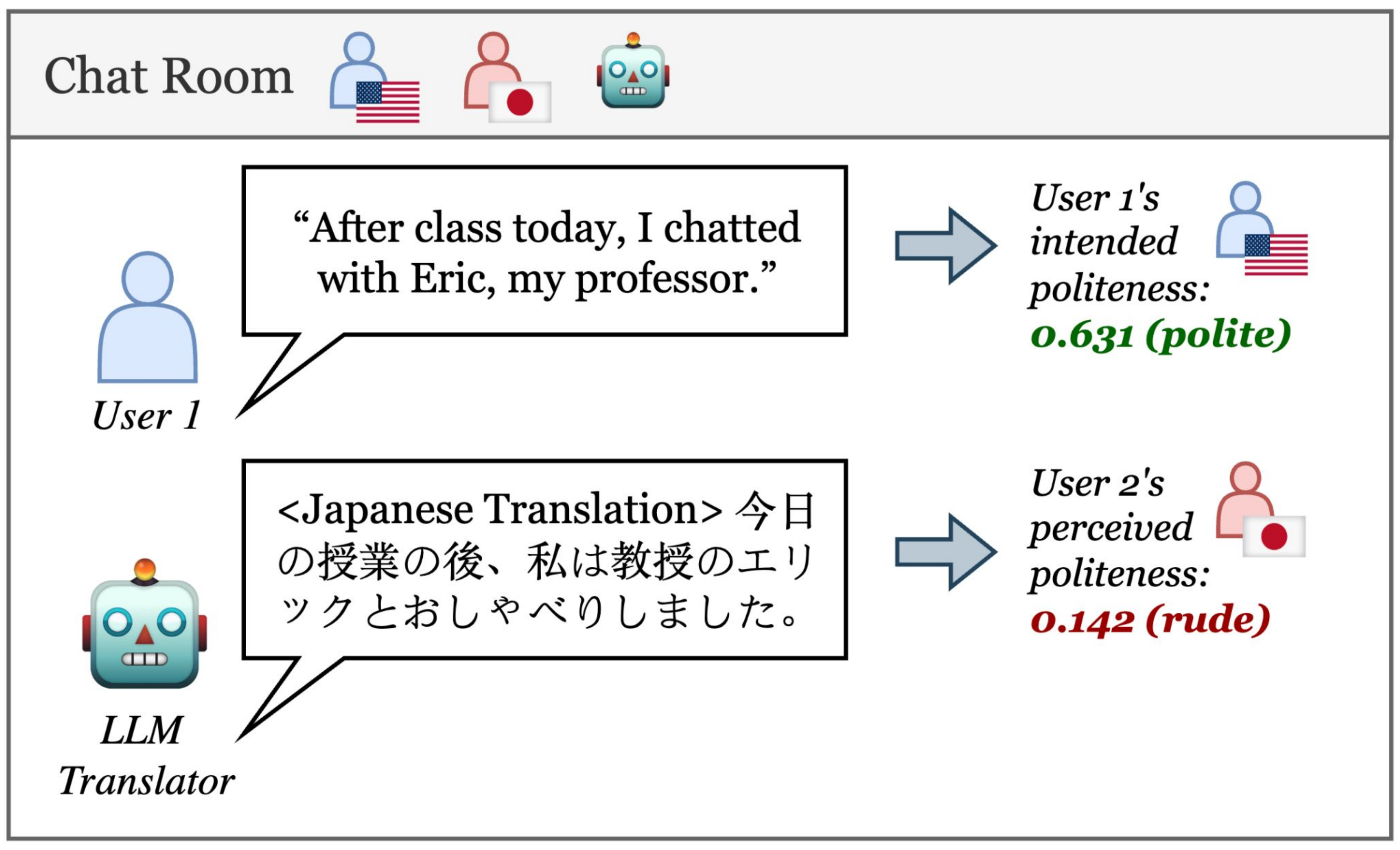
Towards Style Alignment in Cross-Cultural Translation

Shreya Havaldar*, Adam Stein*, Eric Wong, Lyle Ungar

Penn Engineering

ACL 2025 VIENNA
JULY 27 - AUGUST 1

Culture influences appropriate style


Chat Room

User 1

"After class today, I chatted with Eric, my professor."

User 1's intended politeness: **0.631 (polite)**

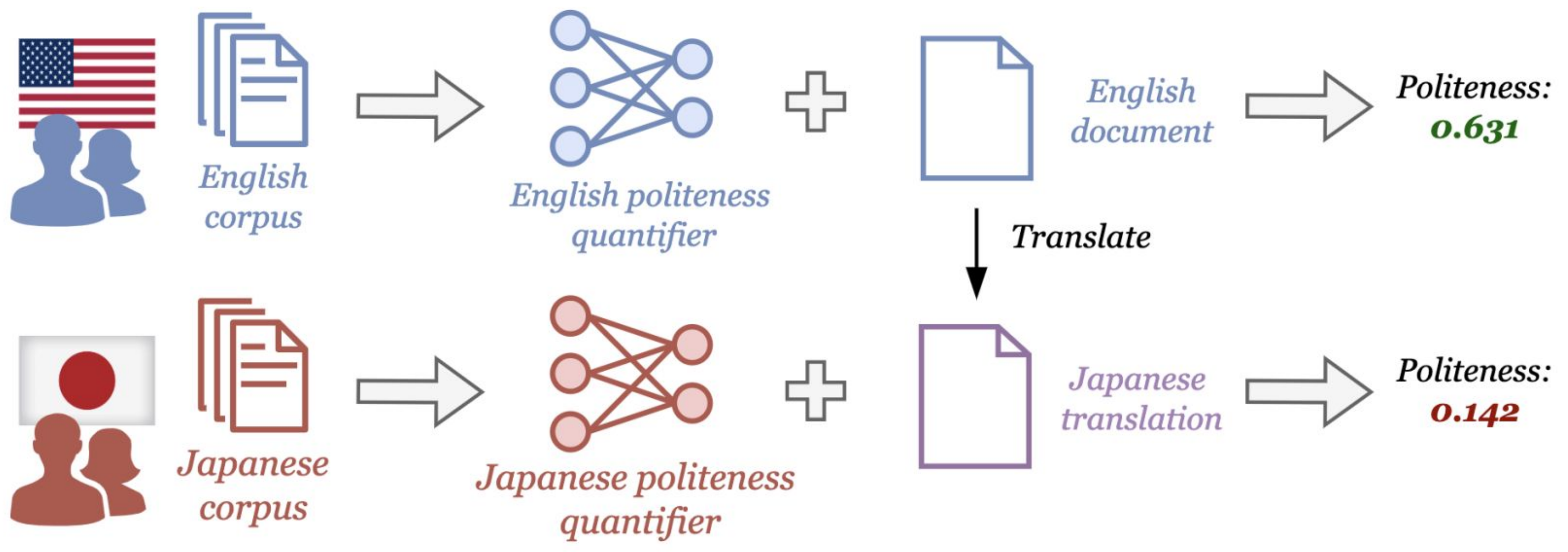
LLM Translator

<Japanese Translation> 今日の授業の後、私は教授のエリックとおしゃべりしました。

User 2's perceived politeness: **0.142 (rude)**

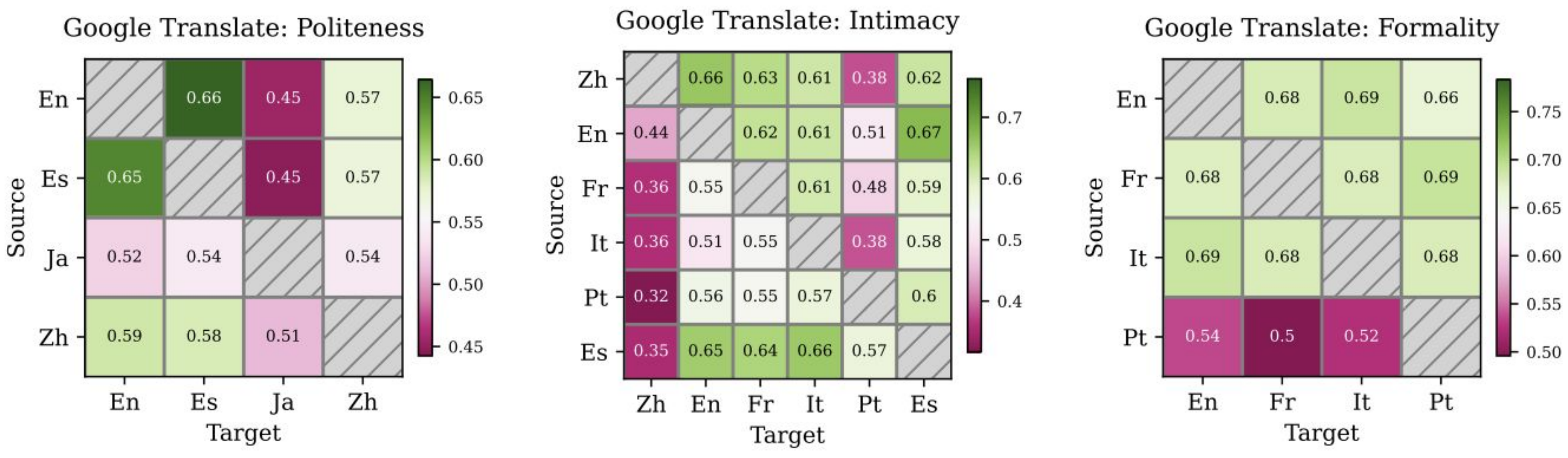
Successful communication depends on the **speaker's intended style** (what the speaker is trying to convey) aligning with the **listener's interpreted style** (what the listener perceives)

Evaluating Style Preservation in LLMs

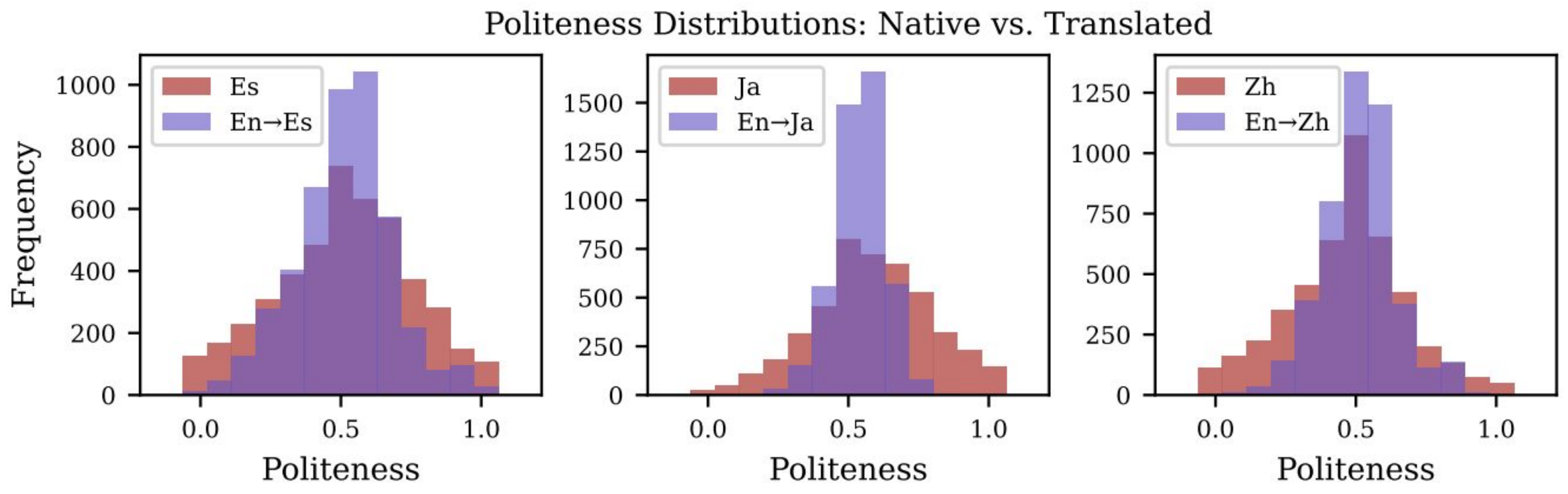


Building a style alignment metric: (1) We select a multilingual corpus X annotated for style by native speakers. (2) We train style quantifiers C_1 and C_2 to label style in L_1 and L_2. (3) Using these quantifiers, we can measure style alignment A using:

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

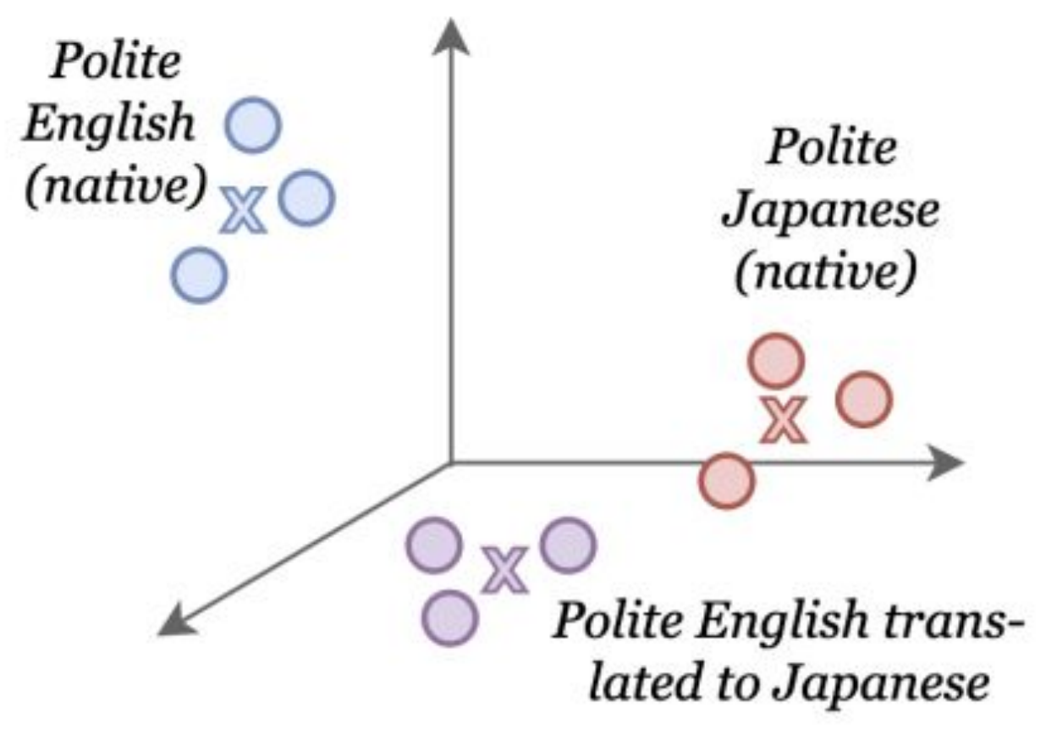


Failure 1: LLMs perform worst in non-Western languages.
Green indicates above average $A(L_1, L_2)$; pink indicates below

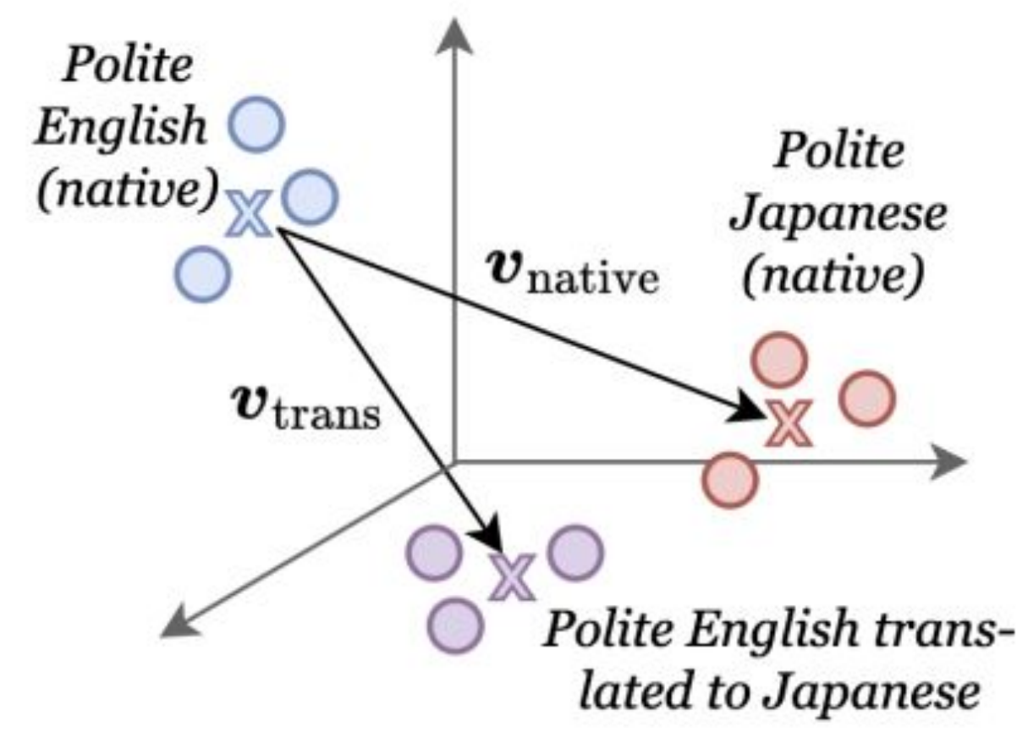


Failure 2: LLMs bias translations towards neutral, reducing real-world variance.

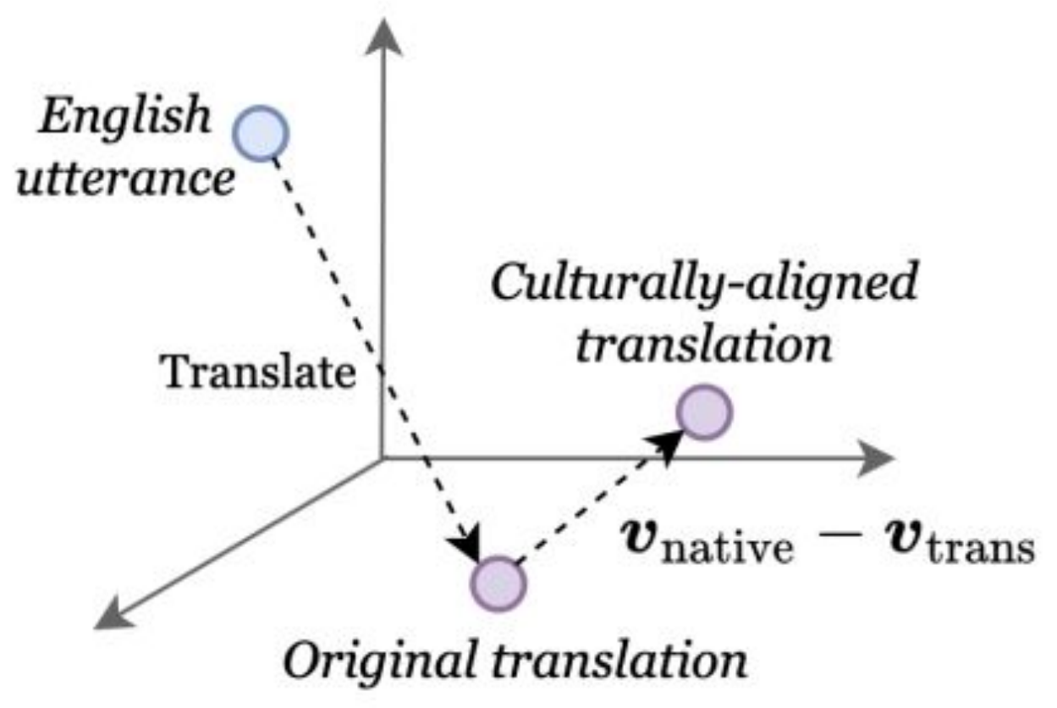
Retrieval-Augmented STyle Alignment



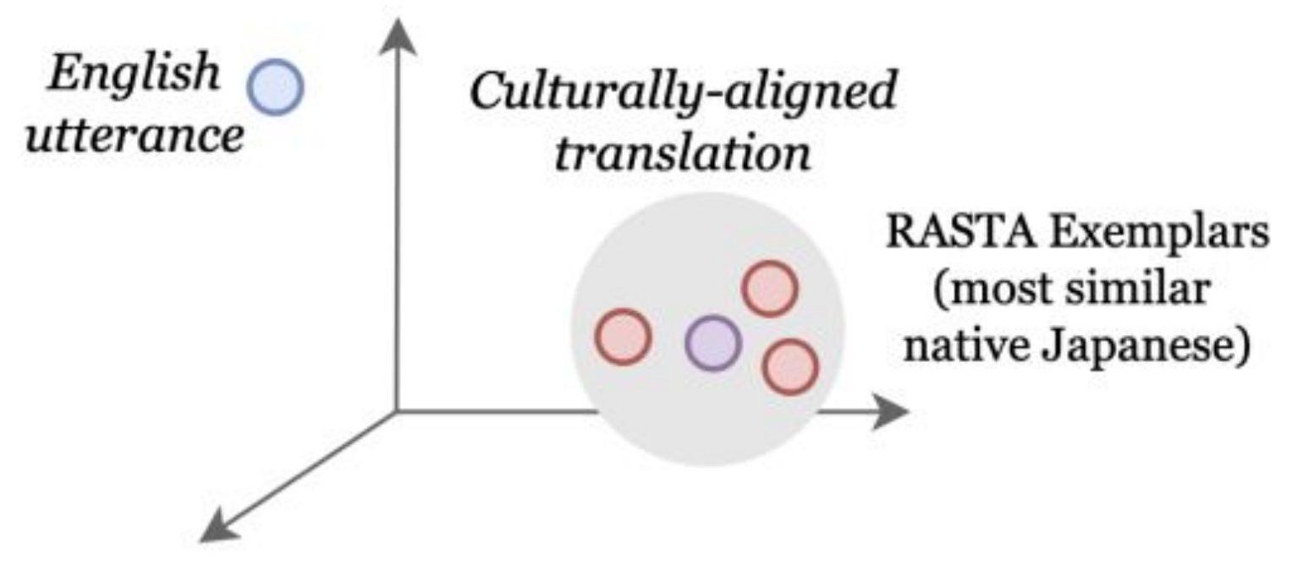
Step 1: Embed data & discover stylistic concepts in embedding space *(e.g. polite Japanese)*



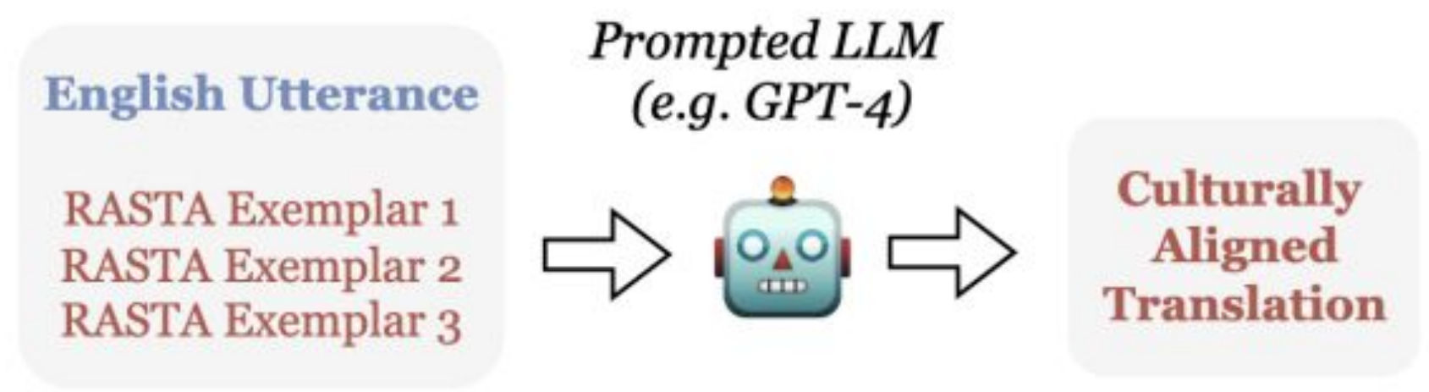
Step 2: Learn mappings between stylistic concepts *(e.g. native polite English → translated polite Japanese)*


English utterance
Translate
Culturally-aligned translation
$v_{\text{native}} - v_{\text{trans}}$
Original translation

Step 3: Apply learned mapping $v_{native} - v_{trans}$ to align input embedding according to cultural communication norms


English utterance
Culturally-aligned translation
RASTA Exemplars (most similar native Japanese)

Step 4: Use the aligned embedding to select the best native few-shot exemplars



Step 5: Generate a more culturally aligned translation

Note: We do this for every style and language pair, as centroids and exemplars are unique to style level and language.

RASTA Improves Translation

Politeness results across English, Spanish, Chinese, & Japanese:

Translation Technique	Style Alignment	Comet-Kiwi	GEMBA
Vanilla	0.53	0.78	95.18
"Preserve Style" Prompting	0.60	0.78	95.56
RASTA (ours)	0.70	0.77	95.13
Average Δ	*+ 24.4%*	*- 1.3%*	*-0.2%*

Overall, RASTA improves translation by:

1. **Significantly increasing style alignment** without degrading translation quality.
2. De-biasing translation performance by **improving alignment in non-Western languages.**
3. **Preserving native speaker variance** and generating translations **preferred by humans** on a Prolific study.

Our paper

My website

Code/Data