Semantic Label Drift in Cross-Cultural Translation

Mohsinul Kabir[†‡], Tasnim Ahmed[♣], Md Mezbaur Rahman[♠], Polydoros Giannouris[†], Sophia Ananiadou[†‡¶]

[†]Department of Computer Science, National Center for Text Mining, The University of Manchester, [‡]ELLIS Manchester

[¶]Aristotle University of Thessaloniki, Archimedes/Athena Research Center, Greece

[♣]School of Computing, Queen's University, Ontario, Canada

[♠]Computer Science, University of Illinois Chicago

{mdmohsinul.kabir, sophia.ananiadou, polydoros.giannouris}@manchester.ac.uk, tasnim.ahmed@queensu.ca, mrahma56@uic.edu

Abstract

Machine Translation (MT) is widely employed to address resource scarcity in low-resource languages by translating data from high-resource languages. While sentiment preservation in translation has long been studied, a critical but underexplored factor is the role of cultural alignment between source and target languages. In this paper, we hypothesize that semantic labels drift or are altered during MT due to cultural divergence. Through a series of experiments across culturally sensitive and neutral domains, we establish three key findings: (1) MT systems, including modern Large Language Models (LLMs), induce label drift during translation, particularly in culturally sensitive domains; (2) unlike earlier statistical MT tools, LLMs encode cultural knowledge, and leveraging this knowledge can amplify label drift; and (3) cultural similarity or dissimilarity between source and target languages is a crucial determinant of label preservation. Our findings highlight that neglecting cultural factors in MT not only undermines label fidelity but also risks misinterpretation and cultural conflict in downstream applications. We release our codebase to facilitate future research in cross-cultural translation: https://github.com/mohsinulkabir14/label_drift. *This paper includes examples that may contain offensive or sensitive language, presented solely for research and demonstration purposes.*

1. Introduction

Machine Translation (MT) has long served as a vital technology for enabling cross-cultural communication, with applications in chat systems, customer support, and social media. Beyond direct communication, MT plays a crucial role in research by facilitating dataset reuse across languages, particularly in low-resource settings. For many low-resource languages in regions such as South Asia and Africa, translating English datasets has become a common strategy to support NLP development (Steigerwald et al., 2022; Nekoto et al., 2020). This approach allows these linguistic communities to benefit from shared knowledge and mitigate data scarcity. With the emergence of large language models (LLMs), this paradigm has become even more prevalent. Translated data are now widely used for benchmarking tasks and for both pretraining and fine-tuning LLMs across diverse linguistic domains.

A critical concern in such dataset reuse is the preservation of semantic labels during translation. Prior studies (Memon et al., 2021; Mohammad et al., 2016) have shown that translating datasets from high-resource to low-resource languages can yield high label preservation, with only $2-3\%$ degradation. However, these findings apply largely to straightforward sentiment analysis tasks with binary labels (positive/negative), where cultural divergence plays a minimal role. In more nuanced research areas, such as affective computing, subtler linguistic cues become crucial. While LLMs and modern MT systems perform well in translating factual or *propositional* content, they still struggle with *non-propositional* aspects, such as politeness, formality, or emotion. Studies by Mirkin et al. (2015) and Rabinovich et al. (2016) reveal that translation can obscure socio-demographic cues like gender and personality traits, while Troiano et al. (2020) demonstrate that transformer-based NMT systems often lose emotional content during translation. Similarly, Havaldar et al. (2025) show that cultural differences can cause misalignment between a speaker's intended style and a listener's interpretation; for example, politeness is often lost in translation.

Translating text into culturally sensitive domains, such as mental health discourse, sarcasm, or irony, requires more than literal fidelity. It demands the preservation of affective and stylistic elements, such as *empathy* and *tone*, to bridge cultural gaps effectively. Despite this, research remains limited on how translation may alter emotional polarity or even cause label drift between source and target languages, including when translations are generated by state-of-the-art (SOTA) LLMs. In this study, we explore the hypothesis that semantic label drift predominantly occurs in affective and culturally sensitive contexts, systematically examining failures in label preservation during cross-cultural translation. Our experiments include both traditional statistical MT systems and contemporary LLMs as translation

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3714–3724
11–16 May 2026.

tools. Specifically, we address the following four research questions:

RQ1: Do state-of-the-art machine translation systems (e.g., Google Translate, NLLB, and modern LLMs) alter dataset labels in culturally sensitive domains during translation? (Yes)

RQ2: Does the cultural knowledge of LLMs reduce or amplify label drift between source and target languages? (May amplify)

RQ3: Does cultural similarity between source and target languages mitigate label drift? (Yes, but in specific domains)

RQ4: Is label drift in translated texts culture-specific, or does it also occur in culturally neutral domains? (Mostly culture-specific)

In addition to analyzing label preservation quantitatively, we also perform a qualitative analysis of translated texts, identifying critical cases of translation refusal by LLMs and discussing instances of cultural misalignment that may result in culturally inappropriate or distorted translations.

2. Related Works

Statistical Machine Translation and Sentiment Alteration. Statistical Machine Translation (SMT) systems such as Google Translate remain widely used and effective for high-resource languages. However, true human-like translation requires a deeper understanding of semantics and context (Weaver, 1952), which SMT and even Neural Machine Translation (NMT) models often fail to capture, particularly regarding emotional nuance and cultural subtleties. Several studies have shown that MT can alter sentiment polarity. For example, Salameh et al. (2015) report that Arabic social media posts translated into English frequently lose or neutralize their original sentiment. Similarly, Saadany et al. (2023) find that online NMT systems often invert or erase emotional tone in tweets. To address this, Kumari et al. (2021) propose fine-tuning global-attention NMT models using actor–critic reinforcement learning with sentiment- and semantics-based rewards, improving preservation in low-resource translation.

Cultural Codes of Language. Although basic human behaviors are universal, their linguistic expressions are deeply culture-specific. Each language encodes norms, prohibitions, and imperatives differently. From this perspective, Aydarova et al. (2024) analyze the axiological (value-related) and cultural codes embedded in behavioral verbs across Russian, Tatar, and English, revealing both shared and unique semantic patterns. They argue that overlooking such codes can lead to cultural misinterpretation in MT. Troiano et al. (2020) empirically demonstrate that MT systems systematically degrade non-propositional emotional information and propose a re-ranking method to mitigate this loss. More recently, Havaldar et al. (2025) show that effective cross-cultural translation is essential to convey affective styles such as politeness or intimacy, as these are shaped by cultural norms. Their findings indicate that even modern LLMs frequently fail to preserve such stylistic cues, often neutralizing or misrendering them, particularly in non-Western languages.

3. Evaluating Label Preservation in Cross-Cultural MT

3.1. Culturally Sensitive Domains

We hypothesize that label alteration occurs primarily in culturally sensitive domains. To select appropriate domains, we consider two factors: cultural sensitivity established in prior literature and availability of datasets. Based on these, we focus on mental health and irony. Mental health concepts, diagnoses, and treatments are shaped by cultural beliefs and practices (Rai et al., 2024; Adebayo et al., 2024), with numerous studies reporting stigma across individualistic and collectivist societies and the risks of cultural insensitivity in diagnosis, assessment, or intervention (Kirmayer, 1989; Altweck et al., 2015; Lyons et al., 2025). Likewise, irony is present across languages but varies in definition, function, and recognition across cultures (Weizman, 2022), shifting from explicit to subtle forms between individualistic and collectivist contexts (Ervas and Schnell, 2024).

To analyze label drift, we require granular datasets in these two domains with multi-class annotations, primarily based on data collected from *Western* users. Accordingly, we select the following datasets:

1. **DEPTWEET** (Kabir et al., 2023a): tweets labeled into four depressive categories (*Non-depressed*, *Mildly depressed*, *Moderately depressed*, *Severely depressed*).

2. **SemEval-2018 Task 3** (Van Hee, 2017): tweets labeled into four irony categories (*Non-ironic*, *Irony by clash*, *Situational irony*, *Other irony*).

We randomly select 1150 samples from each dataset for our experiments, ensuring an equal number of instances from each class. However, the *Situational irony* and *Other irony* classes in the original **SemEval-2018 Task 3** dataset contain only about 200 samples, so we include all available instances from these classes.

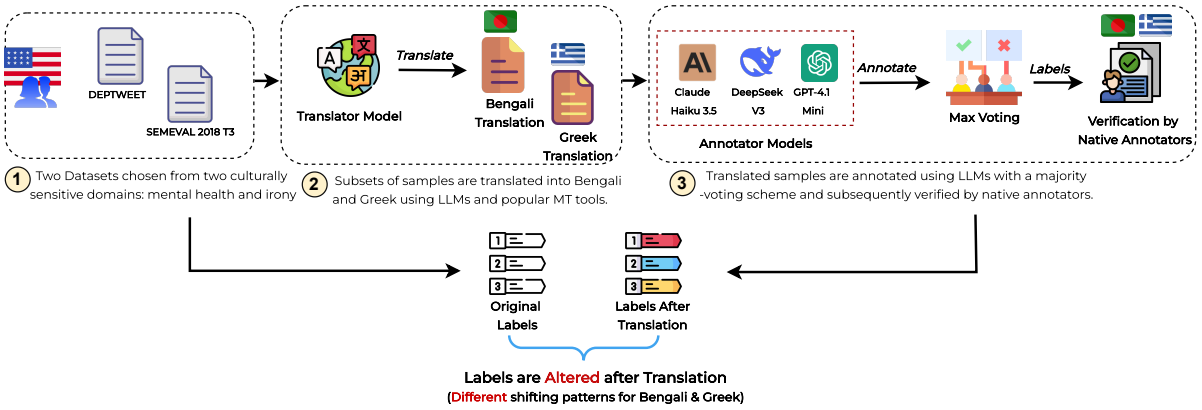


Figure 1: Evaluating label drift in cross-cultural translation. We select two culturally sensitive datasets constructed in Western contexts, translate subsets of randomly sampled data into Bengali and Greek, and annotate them using LLMs with verification by native annotators. By comparing translated annotations with the original labels, we assess the extent of label preservation during translation.

3.2. Models & Probing Methods

We aim to simulate the experience of an average user relying on online translation tools. To this end, we evaluate a mix of translation systems alongside both open-source and proprietary LLMs. Specifically, our experiments include Google Translate, NLLB-1.3B, GPT-4.1 Mini, Claude Sonnet 4, Llama-3.3 (70B), and DeepSeek V3.

> Imagine you are a {marital_status} {sex} from {region}, {country}. You are {age} years of age and completed {education} education level. You consider yourself part of the {social_class}. Translate the following English sentence into {target_language} from this perspective. {Sentence:}

Figure 2: Translation with Anthropological Prompting.

For LLMs, we design two prompting strategies. In the first, we mimic typical user behavior by issuing a simple instruction: *"Provide translation of the following sentence from English to {target_language}."* Translations generated this way are referred to as *Literal Translation*.

In the second, we leverage models' cultural knowledge through *Anthropological Prompting* (AlKhamissi et al., 2024), which grounds the instructions in anthropological contexts by asking the model to assume a culturally situated perspective. The method uses six demographic dimensions, and the authors empirically show that optimal cultural alignment with the respective culture occurs with the following specific attributes: `Region`: Country-Specific, `Sex`: Male, `Age`: < 50, `Social Class`: Upper/Lower Middle Class, `Education Level`: Higher, and `Marital Status`: Married. We adopt this setup to maximize cultural alignment, and translations generated in this way are referred to as *Cultural Prompting* (see Figure 2).

3.3. Country & Language Selection

Although using language as a direct proxy for culture has been criticized (Kabir et al., 2025), it remains unavoidable in the context of translation tasks. We hypothesize that when the source and target languages belong to culturally similar contexts, dataset labels are better preserved, with fewer alterations. Since both datasets in our study originate primarily from North-American and European user data, we select one culturally similar language and one contrasting language. For cultural comparison, we rely on the Inglehart-Welzel World Cultural Map (World Values Survey Association, 2023), which is structured around two dimensions: Traditional vs. Secular-rational values and Survival vs. Self-expression values. We also prioritize low-resource languages, as translation is especially relevant for such domains. Based on these criteria, we choose **Greek**, representing an individualistic culture similar to the Anglo-centric world, and **Bengali**, representing a contrasting collectivist culture.

3.4. Translation & Annotation

We adopt a Human–LLM collaboration scheme (Wang et al., 2024) to annotate translated samples. For consistency, we follow the annotation schemes described in the original DEPTWEET (Kabir et al., 2023a) and SemEval-2018 T3 (Van Hee, 2017) papers, where authors detail their annotation procedures. These schemes are converted into detailed annotation guidelines, which we then use as prompting instructions for LLM-based annotation.

DEPTWEET Dataset (Mental Health)									
Bengali Translation									
	Literal Prompting				*Cultural Prompting*				Δ(Lit.-Cul.)
	Non-Depressed	**Mild**	**Moderate**	**Severe**	**Non-Depressed**	**Mild**	**Moderate**	**Severe**	
Google Translate	44.3	50.7	58.3	84.7	-	-	-	-	-
NLLB-1.3B	**56.4**	53.1	54.9	84.3	-	-	-	-	-
Claude-4-Sonnet	47.7	46.9	**62.2**	82.6	48.4	48.6	**63.5**	76.0	+0.73
GPT-4.1-Mini	43.9	52.4	61.1	86.1	44.6	56.2	60.1	78.7	+0.98
Llama-3.3	48.4	49.0	53.5	**87.8**	48.1	48.6	53.5	**87.5**	+0.25
DeepSeek V3	54.0	**58.7**	55.9	84.3	**54.7**	**58.7**	58.0	82.2	-0.18
Greek Translation									
	Literal Prompting				*Cultural Prompting*				Δ(Lit.-Cul.)
	Non-Depressed	**Mild**	**Moderate**	**Severe**	**Non-Depressed**	**Mild**	**Moderate**	**Severe**	
Google Translate	40.4	51.4	57.6	84.0	-	-	-	-	-
NLLB-1.3B	**57.1**	**56.9**	55.6	77.7	-	-	-	-	-
Claude-4-Sonnet	48.1	48.3	61.8	86.1	49.5	49.0	**62.2**	81.9	+0.43
GPT-4.1-Mini	48.4	50.3	**62.2**	86.8	48.8	52.1	60.8	79.1	+1.73
Llama-3.3	42.2	47.9	54.9	**88.9**	42.9	46.2	52.1	**85.4**	+1.83
DeepSeek V3	54.0	52.8	60.1	85.7	**53.7**	**52.4**	60.4	85.4	+0.18
Δ ***(Greek-Bengali)***$_{deptweet}$	-0.75	-0.53	+1.05	-0.10	-0.23	-3.10	+0.10	+1.85	
SemEval-2018 T3 Dataset (Irony)									
Bengali Translation									
	Literal Prompting				*Cultural Prompting*				Δ(Lit.-Cul.)
	Non-ironic	**Ironic-clash**	**Situational-irony**	**Other-irony**	**Non-ironic**	**Ironic-clash**	**Situational-irony**	**Other-irony**	
Google Translate	29.7	**81.8**	21.0	0.5	-	-	-	-	-
NLLB-1.3B	53.1	56.7	**31.3**	**3.0**	-	-	-	-	-
Claude-4-Sonnet	48.9	65.3	25.2	2.5	53.4	**64.7**	19.2	0.0	+1.15
GPT-4.1-Mini	51.4	67.1	22.0	2.5	54.4	62.6	22.0	**3.0**	+0.25
Llama-3.3	**62.2**	47.2	20.6	2.0	55.4	54.6	21.0	2.5	-0.38
DeepSeek V3	58.4	62.3	28.5	1.0	**58.4**	62.3	**28.0**	1.0	+0.13
Greek Translation									
	Literal Prompting				*Cultural Prompting*				Δ(Lit.-Cul.)
	Non-ironic	**Ironic-clash**	**Situational-irony**	**Other-irony**	**Non-ironic**	**Ironic-clash**	**Situational-irony**	**Other-irony**	
Google Translate	32.1	**78.3**	21.5	2.5	-	-	-	-	-
NLLB-1.3B	51.6	59.3	**33.2**	3.5	-	-	-	-	-
Claude-4-Sonnet	58.4	69.4	28.0	3.0	**62.4**	64.4	25.7	4.0	+0.58
GPT-4.1-Mini	59.1	71.2	29.9	3.5	61.4	67.7	**29.9**	3.5	+0.30
Llama-3.3	**64.2**	57.6	21.0	**4.0**	55.4	60.2	18.7	3.0	+2.38
DeepSeek V3	57.1	71.8	25.7	4.0	56.1	**73.0**	24.8	**4.0**	+0.18
Δ ***(Greek-Bengali)***$_{semeval2018}$	+3.13	+4.53*	+1.78	+1.50*	+3.43	+5.28	+2.23	+2.00	

Table 1: Label preservation rates (%) for the **Mental Health** (top) and **Irony** (bottom) datasets translated into Bengali and Greek using six MT models. Results are shown for both *Literal* and *Cultural* prompting. **Bold** numbers indicate the highest preservation per class. Mean differences between Greek and Bengali (Δ(Greek–Bengali)) and between *Literal* and *Cultural* prompting (Δ (Lit.-Cul.)) are reported. Statistical significance is assessed via one-sided Wilcoxon signed-rank tests ($^*p < 0.05$). No significant gain is observed for cultural over literal prompting; only *Ironic-clash* and *Other-irony* show significant cross-lingual differences.

LLMs have been shown to produce high-quality annotations in diverse NLP tasks, often rivaling or surpassing crowdsourced annotators (He et al., 2024; Tan et al., 2024). However, given the complexity of mental health and irony, which require nuanced interpretation, we adopt the following strategy:

1. **Majority Voting:** Each sample is annotated by three LLMs: GPT-4.1 Mini, Claude Haiku 3.5, and DeepSeek V3, selected for their cost and speed-efficiency. The final label is determined via majority voting.

2. **Human Validation:** Native speakers of Bengali and Greek review ambiguous cases (e.g., instances without consensus such as three distinct labels). They also annotate a subset of translated samples to validate the LLM-derived majority-vote labels.

This Human–LLM collaboration approach balances cost-effectiveness with reliability in complex domains that is best suited for our study. (Wang et al., 2024). Our full experiment pipeline is demonstrated in Figure 1.

3.5. Evaluation Metrics

To assess the consistency between the original and translated labels, we employ three complementary evaluation metrics:

- **Label Preservation Rate**: Measures the proportion of labels that remain unchanged after translation.

- **Kullback–Leibler (KL) Divergence**: Captures the overall distributional shift of labels between the source and translated datasets.

- **Matthews Correlation Coefficient (MCC)**: Evaluates the strength of correspondence between original and translated labels.

4. Findings

RQ1. Translation Preserves Perceived Severity but Obscures Mild Presentations

Table 1 presents the label preservation rates following translation for both datasets. A clear pattern emerges where labels for the extreme classes are consistently well-preserved. For the **Mental Health** dataset, the *severe* depression category exhibits high preservation rates in both Bengali and Greek. However, the translation process introduces a systematic bias towards inflating perceived severity, evidenced by the considerably lower preservation rates for the *mild* and *moderate* classes. This indicates that *non-severe* cases are frequently drifted to more *severe* categories after translation.

This finding is further quantified by calculating the Matthews Correlation Coefficient (MCC) to measure the agreement between original and post-translation labels. As shown in Figure 3, strong agreement is observed for the *Non-Depressed* and *Severe* classes. In contrast, the *Mild* and *Moderate* classes show weak agreement, with much lower MCC scores. This confirms that the nuanced linguistic distinctions between mild and moderate depression are largely lost or altered during translation, making reliable classification of these categories after translation exceedingly difficult.

A similar but distinct pattern is observed for the **Irony** dataset. The *non-ironic* and *ironic-clash* categories are well-preserved across languages, suggesting that translation robustly preserves non-irony and obvious, clash-based irony. Conversely, the more context-dependent *situational irony* is highly degraded. The MCC scores tell a more comprehensive story, revealing weak to moderate agreement for all irony categories, indicating that ironic content is generally not robust to machine translation. Most notably, the *other-irony* category frequently yields MCC scores near or below zero, demonstrating that its classification after translation becomes effectively unpredictable and uncorrelated with the original labels.

To understand the macro-level effect of translation on the entire dataset, we report the Kullback–Leibler (KL) Divergence between the original and translated label distributions in Figure 4. For the **Mental Health dataset**, we observe low KL Divergence scores (all < 0.08). This indicates that while individual instances are altered (as shown by low preservation and MCC for mild/moderate classes), the global distribution of severity labels remains largely intact. This suggests a systematic re-categorization within the severity spectrum rather than a wholesale shift. In stark contrast, the **Irony** dataset exhibits very high KL Divergence (up to 0.65). This confirms that translation induces a severe distortion of the entire label distribution, fundamentally altering the proportion of *ironic* to *non-ironic* content and between different irony types. This macro-level distortion complements our instance-level findings, illustrating that the translation process has a catastrophic effect on the fabric of ironic discourse.

RQ2. Cultural Knowledge in LLMs may Exacerbate Label Drift during Translation

We ground the LLM responses in relative cultural context by using the *anthropological* prompting strategy shown in Figure 2 to answer RQ2. Initial results in Table 1 show a slightly higher label preservation rate for literal prompting compared to cultural prompting, though this difference is not statistically significant.

Dataset	ΔMCC (Lit.-Cul.)	p-value (wilcoxon)
DT-Bengali	0.0044	0.2589
DT-Greek	0.0142*	0.0035
SE T3-Bengali	0.0041	0.5921
SE T3-Greek	0.0112*	0.0207
Overall	0.0085*	0.0026

Table 2: Mean MCC differences (Literal - Cultural) for DEPTWEET (DT) and SemEval-2018 T3 (SE T3) datasets. Positive values indicate superior performance of literal prompting. Statistical significance is assessed with a Wilcoxon signed-rank test (* $p < 0.05$).

To quantify this effect, we perform a paired t-test on the Matthew's Correlation Coefficient (MCC) scores to compare the correlation of literal and cultural prompting with the original labels. A Wilcoxon signed-rank test (Table 2) reveals a small but statistically significant overall advantage for literal prompting over cultural prompting (Mean $\Delta\text{MCC}_{\text{Literal - Cultural}} = +0.0085$, 95% CI [0.0039, 0.0132], $p < 0.01$).

This overall effect is primarily driven by Greek translations, where literal prompting demonstrates significantly stronger agreement with original labels on both mental health ($\Delta\text{MCC} = +0.0142$, $p < 0.01$) and irony ($\Delta\text{MCC} = +0.0112$, $p = 0.02$) tasks. In contrast, for Bengali translations, we find no significant difference between prompting strategies for either task (both $p > 0.13$).

Based on these observations, we conclude that explicitly instructing models to consider cultural context, while intended to improve relevance, can instead motivate label drift by altering the textual features critical for classification. The superior per-

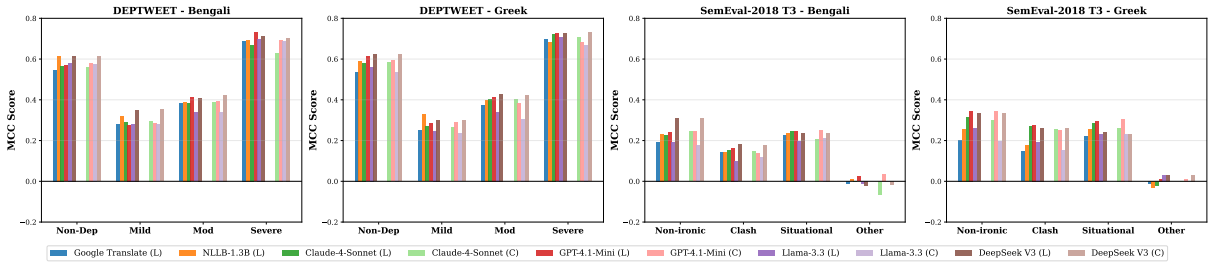


Figure 3: Matthews Correlation Coefficient (MCC) scores comparing the performance of six translation/language models. The grouped bar charts display literal prompting (L) results (darker bars) for all models, while cultural prompting (C) results (lighter bars) are shown only for the four LLM-based models. MCC scores interpret as: $0.0 - 0.3$:weak, $0.31 - 0.5$: weak moderate, $0.51 - 0.7$: moderate strong, and > 0.7 : strong performance.

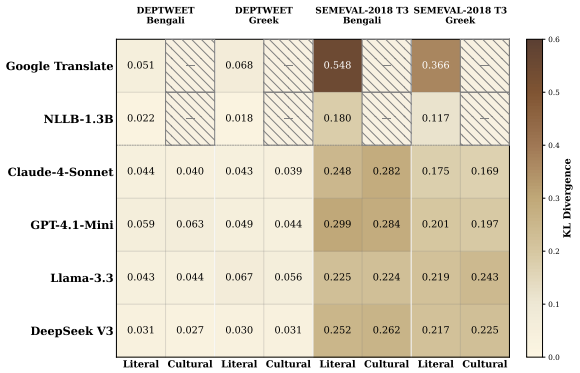


Figure 4: Kullback–Leibler (KL) divergence between original and translated label distributions across six MT models under different prompting techniques. The SemEval-2018 T3 (irony) dataset exhibits greater divergence from the original labels than DEPTWEET.

formance of literal prompting suggests that for the specific task of preserving pre-defined labels, a direct translation approach is unexpectedly more reliable and robust.

RQ3: Cultural Similarity between Source and Target Languages can Mitigate Label Drift in Specific Domains

To investigate the role of cultural similarity, we select language pairs with varying proximity to Anglo-American culture: Greek (a European language with strong historical ties) and Bengali (a South Asian language with a distinct cultural context). We hypothesize that translations into culturally closer languages would experience less label drift.

Our quantitative results provide strong but domain-specific support for this hypothesis. Table 1 reports the mean difference in label preservation between Greek and Bengali translations for the mental health (DEPTWEET) and irony (SemEval-2018 T3) datasets. No statistically significant differences emerge in the mental health dataset, whereas the irony dataset shows significantly better preservation in the *Ironic-clash* and *Other-irony* categories. A consistent pattern appears in Table 3, which presents differences in Matthew's Correlation Coefficient (MCC), where positive values indicate stronger label correlation for Greek than Bengali. In the mental health domain, MCC differences are negligible under both literal ($\Delta = +0.001$) and cultural ($\Delta = -0.005$) prompting, with no significant effects at the category level. By contrast, in the irony domain, Greek translations consistently achieve higher label correlation. Under literal prompting, the average MCC is statitically higher (Δ = $+0.039$, $p < 0.05$), with significant gains in the *Non-Ironic*, *Ironic-Clash*, and *Situational-Irony* categories. Cultural prompting shows a similar, though not statistically significant, trend ($\Delta = +0.048$).

This divergence in outcomes between irony and

	△MCC(Greek - Bengali)	
	Literal	**Cultural**
DEPTWEET		
Non-Depressed	0.002	0.002
Mild	-0.019	-0.032
Moderate	0.009	-0.008
Severe	0.012	0.017
Average	0.001	-0.005
SemEval-2018 T3		
Non-Ironic	0.055*	0.050
Ironic Clash	0.074*	0.087
Situational	0.023*	0.030
Other	0.003	0.025
Average	0.039*	0.048

Table 3: Mean MCC differences for (Greek - Bengali) across dataset categories. Positive values indicate better correlation of Greek translation with original labels. Statistical significance assessed via one-tailed Wilcoxon signed-rank test (* $p < 0.05$).

mental health motivates a qualitative analysis to uncover underlying mechanisms. In irony, the relative success of literal prompting in Greek likely reflects cultural proximity to Anglo-American humor, where ironic intent often survives direct translation. Conversely, in the mental health domain, native speaker feedback reveals that cultural prompting can amplify emotional tone (e.g., translating "I feel uneasy" as "this is driving me crazy") or censor explicit references to self-harm (e.g., replacing "suicide" with "dark thoughts"). These adaptations either intensify or neutralize emotional severity, leading to label drift. Overall, while cultural prompting enhances cultural naturalness through idiomatic expression, it tends to increase semantic divergence, whereas literal prompting maintains closer alignment with the source labels.

In conclusion, cultural similarity between source and target languages can reduce label drift, but its effect is domain-dependent. For conceptually aligned domains like irony, direct translation into a culturally proximal language such as Greek may be highly effective in label preservation. However, for culturally sensitive domains like mental health, cultural adaptation may introduce an additional layer of shift, offsetting the potential benefits of cultural proximity.

RQ4: Label Drift Predominantly Occurs in Culturally Sensitive Domains

We investigate whether the phenomenon of label drift extends across domains regardless of cultural sensitivity. Prior studies have reported that translations of relatively simple domains, such as positive/negative sentiment, can achieve competitive accuracy with minimal label alteration (Mohammad et al., 2016; Memon et al., 2021). To build on this, we select a domain that is expected to involve little to no cultural sensitivity, i.e., culturally agnostic emotions. For dataset selection, we consider two criteria: (i) the dataset should contain granular labels, and (ii) the samples should be sufficiently long to make annotation appropriately challenging and complex. Based on these criteria, we choose the Amazon Product Review dataset (Hou et al., 2024), which consists of lengthy real-world product reviews accompanied by ratings from 1 to 5. To facilitate classification, we group the lowest scores (1 and 2) as *Negative (Neg.)*, the middle score (3) as *Neutral (Neu.)*, and the highest scores (4 and 5) as *Positive (Pos.)*, following prior work by Kabir et al. (2023b); Wang et al. (2020), etc. From this dataset, we randomly sample 200 reviews per category and apply the translation and annotation procedure described in Section 3.4. The average sentence length for the selected subset is 88.82 words, making the annotation process considerably challenging for the annotator models.

Language	Model	*Literal Prompting*			*Cultural Prompting*		
		Neg.	Neu.	Pos.	Neg.	Neu.	Pos.
Bengali	Google Translate	0.910	0.460	0.895	–	–	–
	NLLB-1.3B	0.905	0.460	0.910	–	–	–
	Claude-4-Sonnet	0.905	0.475	0.900	0.905	**0.480**	0.905
	GPT-4.1-Mini	0.915	**0.480**	0.905	0.910	0.475	0.915
	Llama-3.3	0.910	0.470	**0.915**	**0.920**	0.465	**0.920**
	DeepSeek V3	**0.915**	0.445	0.910	0.905	0.445	0.910
Greek	Google Translate	0.935	0.395	**0.920**	–	–	–
	NLLB-1.3B	0.940	0.465	0.910	–	–	–
	Claude-4-Sonnet	0.945	**0.470**	0.905	0.940	**0.455**	0.900
	GPT-4.1-Mini	0.950	0.440	0.915	0.950	0.435	**0.915**
	Llama-3.3	0.930	0.445	0.910	0.935	0.435	0.905
	DeepSeek V3	**0.960**	0.450	0.900	**0.950**	0.450	0.910

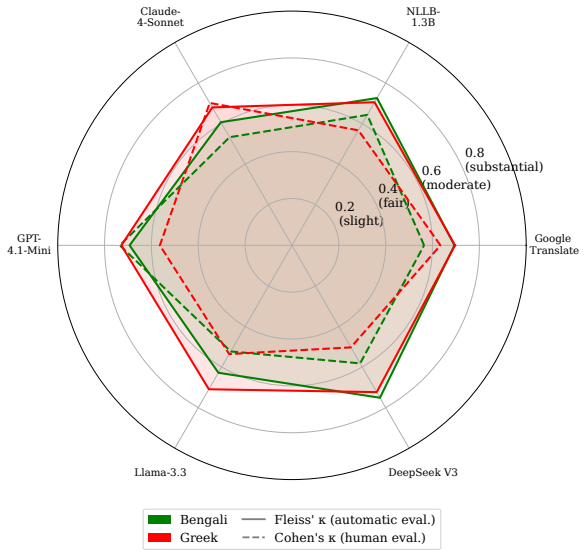
Table 4: Label preservation rates (%) for Amazon Product Review dataset after translation into Bengali and Greek. Bold values indicate the highest preservation per category across models. Label preservation is notably higher than DEPTWEET and SEMEVAL-2018 T3.

Despite these challenges, the results in Table 4 show very high label preservation for the *Negative* and *Positive* samples. The relatively lower preservation rate in the *Neutral* category reflects the well-documented phenomenon of *polarity shift*, where neutral emotions are translated into positive or negative emotions (and vice versa) (Lohar et al., 2017; Hartung et al., 2023). A possible explanation is that models often struggle to assign a neutral label due to shifts in linguistic features introduced during translation (Salameh et al., 2015). Our calculated KL divergence between original and translated labels ranges from 0.046 to 0.053 for Bengali and from 0.061 to 0.072 for Greek, indicating a minimal distributional shift after translation. These findings suggest that label drift is less pronounced in culturally neutral domains, where polarity effects dominate rather than cultural misalignment.

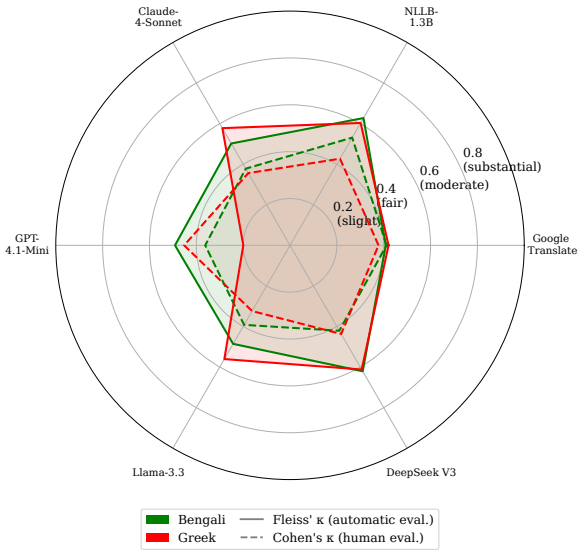
5. Annotation Framework and Agreement Analysis

We employ a majority-voting scheme across three LLMs for automatic annotation. To validate these labels, we randomly select 10% of the data from both DEPTWEET and SemEval-2018 T3 and obtain annotations from two native speakers from each language with prior corpus annotation experience. The human annotators, all PhD students with relevant expertise in linguistics or computational social science in respective languages, followed the same detailed guideline used for automatic annotation.

Inter-Annotator Agreement (IAA) score is measured using Fleiss' Kappa among the three LLMs and Cohen's Kappa between the final automatic labels and human annotations. Figure 5 presents the IAA scores. Automatic annotation achieves moderate to substantial agreement, while agreement between human and automatic labels ranges from fair



(a) IAA Scores for DEPTWEET



(b) IAA Scores for SemEval-2018 T3

Figure 5: Agreement scores for automatic (solid-line) and human (dashed-line) annotations, ranging from fair to substantial agreement across both datasets.

to moderate for DEPTWEET. For SemEval-2018 T3, agreement remains fair to moderate between automatic and human annotations, reflecting the greater difficulty of this dataset.

6. Cross-Cultural Translation: Beyond Label Preservation

6.1. Translation Refusal

We observe that Claude Sonnet 4 and DeepSeek V3 refuse to translate a subset of samples across both Bengali and Greek datasets. Notably, the same sentences are consistently rejected in both languages. We apply HDBSCAN clustering followed by BERTopic (Grootendorst, 2022) modeling to analyze these refused samples, as shown in Figure 6. The rejected content primarily involves sexually explicit material, slang and aggression, gore, and drug-related themes. Further inspection reveals that these sentences belong largely to the *Severely Depressed* class (∼60%) in DEPTWEET and the *Situational Irony* class (∼37%) in SemEval-2028 T3, both minority categories in their respective datasets. Since these samples carry rich but underrepresented information, their exclusion during cross-lingual translation risks losing valuable cultural and linguistic signals, a factor that requires careful consideration in cross-cultural MT.

6.2. Cultural Misalignment

We find a number of samples that contains cultural references in the specific culture, and gets completely lost when translated in Bengali or Greek.

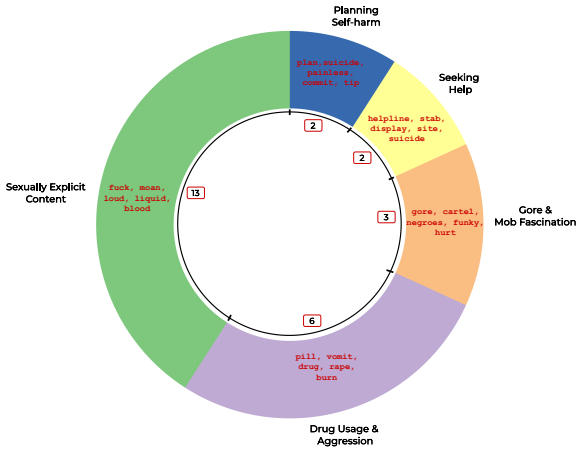
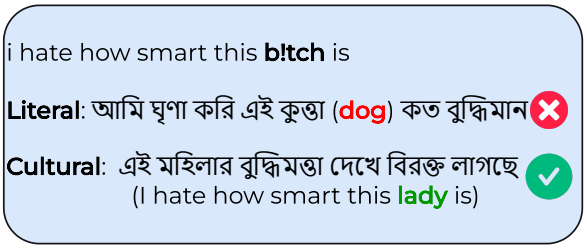


Figure 6: Topic distribution of refused samples showing five identified themes. Outer labels indicate topic categories, keywords inside represent the most characteristic terms, and center values show sample sizes (sentences per topic)

For instance, the phrase *"I'm not a Lohan anymore"* in the above textbox (taken from SemEval-2018 T3) is a cultural reference to actress Lindsay Lohan and her public struggles. A Bengali or Greek reader unfamiliar with this context would struggle to interpret the underlying meaning. The translation fails to convey the meaning of *"I've left my troubled past behind."* Even the cultural prompting misses

this to align with the respective culture.

Another frequent source of cultural misalignment arises from the use of slang. Slang expressions common in Western contexts can become profane or overtly vulgar when translated into more conservative cultures.



For example, the above sample from the DEPTWEET dataset contains the term *"b!tch"*, which in English is often used playfully, combining admiration with mild insult. However, in Bengali, the literal translation renders it as a highly offensive term. The culturally prompted version mitigates this by replacing the slang with a contextually appropriate feminine expression, aligning better with Bengali social norms. In Greek, the literal translation from NLLB-1.3B "σκύλα" (bitch) retains the insult but sounds harsher than intended, while GPT-4.1-Mini's literal rendering "πουτ@ν@" (whore) amplifies vulgarity, losing the original nuance. Conversely, the cultural translation "γυναίκα" (woman) softens the tone excessively, erasing both the playful and pejorative dimensions. These examples illustrate that slang requires careful handling in cross-cultural translation, a challenge that cultural prompting appears to manage more effectively. In summary, cultural adaptation during translation (Singh et al., 2024) has the potential to effectively resolve the issues of refusal and misalignment by maintaining both semantic fidelity and cultural appropriateness.

7. Conclusion

In this study, we investigate semantic label drift in cross-cultural translation. Focusing on two culturally sensitive domains (mental health and irony), we demonstrate that semantic labels often change after translation, both in traditional statistical machine translation systems (e.g., Google Translate) and in modern LLM–based translators. Beyond label drift, we identify a separate, mutually exclusive phenomenon: culturally inappropriate translations due to misalignment between the source and target language's cultures. Our findings highlight the need to revalidate labels and apply cultural adaptation strategies before reusing translated datasets in cross-cultural NLP research.

Limitation

This study focuses on two culturally sensitive domains: mental health and irony, to examine semantic label drift after translation. While we acknowledge that other domains such as sarcasm (Blasko et al., 2021) and humour (Jiang et al., 2019) also exhibit strong cultural dependencies, their inclusion was beyond the scope of this work due to time and resource constraints. We select mental health and irony specifically because of the availability and granularity of existing annotated datasets in these areas.

Similarly, our cross-cultural experiments are conducted using Bengali and Greek as representative target languages. Although incorporating additional cultures would provide a broader perspective, our objective in addressing RQ3- *examining whether cultural similarity between source and target languages mitigates label drift*, require a focused, controlled comparison. Expanding to more cultures could have diluted this comparative analysis. Moreover, access to qualified native annotators for verifying machine-translated labels is a practical constraint; Bengali and Greek offered the most feasible and representative choices under these conditions.

Acknowledgment

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

8. Bibliographical References

Yusuf Olalekan Adebayo, Raphael Ekundayo Adesiyan, Chibuzor Stella Amadi, Oluwaseun Ipede, Lucy Oluebubechi Karakitie, and Kaosara Temitope Adebayo. 2024. Cross-cultural perspectives on mental health: Understanding variations and promoting cultural competence. *World Journal of Advanced Research and Reviews*, 23(01):432–9.

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

Laura Altweck, Tara C Marshall, Nelli Ferenczi, and Katharina Lefringhausen. 2015. Mental health literacy: a cross-cultural approach to knowledge and beliefs about depression, schizophrenia and generalized anxiety disorder. *Frontiers in psychology*, 6:1272.

Alsou Mirzayanovna Aydarova, Albina Anvarovna Bilyalova, and Tamara Ivanovna Zelenina. 2024. Axiological and cultural codes of verbal semantics in the context of machine translation (based on the verbs of the russian, tatar and english languages). *Philology. Theory & Practice*, 17(8):2758–2762.

Dawn G Blasko, Victoria A Kazmerski, and Shariffah Sheik Dawood. 2021. Saying what you don't mean: A cross-cultural study of perceptions of sarcasm. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 75(2):114.

Francesca Ervas and Zsuzsanna Schnell. 2024. Irony across cultures: A contrastive analysis of conceptualizations and social functions. In *Studying verbal irony and sarcasm: Methodological perspectives from communication studies and beyond*, pages 279–302. Springer.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Kai Hartung, Aaricia Herygers, Shubham Vijay Kurlekar, Khabbab Zakaria, Taylan Volkan, Sören Gröttrup, and Munir Georges. 2023. Measuring sentiment bias in machine translation. In *International Conference on Text, Speech, and Dialogue*, pages 82–93. Springer.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025. Towards style alignment in cross-cultural translation. *arXiv preprint arXiv:2507.00216*.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Tonglin Jiang, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in psychology*, 10:123.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in llms. *arXiv preprint arXiv:2502.08045*.

Mohsinul Kabir, Tasnim Ahmed, Md Bakhtiar Hasan, Md Tahmid Rahman Laskar, Tarun Kumar Joarder, Hasan Mahmud, and Kamrul Hasan. 2023a. Deptweet: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139:107503.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023b. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. *arXiv preprint arXiv:2305.06595*.

Laurence J Kirmayer. 1989. Cultural variations in the response to psychiatric disorders and emotional distress. *Social Science & Medicine*, 29(3):327–339.

Divya Kumari, Asif Ekbal, Rejwanul Haque, Pushpak Bhattacharyya, and Andy Way. 2021. Reinforced nmt for sentiment and content preservation in low-resource scenario. *Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–27.

Pintu Lohar, Haithem Afli, and Andy Way. 2017. Maintaining sentiment polarity in translation of user-generated content. *Prague Bulletin of Mathematical Linguistics*, (108):73–84.

Phoebe Lyons, Auden Edwardes, Laura Bladon, and Kathryn M Abel. 2025. Culturally sensitive mental health research: a scoping review. *BMC psychiatry*, 25(1):190.

Abdul Ghafoor Memon, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Abdullah Somro, Rakhi Batra, and Mudasir Ahmad Wani. 2021. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*.

Sunny Rai, Khushi Shelat, Devansh R Jain, Kishen Sivabalan, Young Min Cho, Maitreyi Redkar, Samindara Sawant, Lyle H Ungar, and Sharath Chandra Guntuku. 2024. Cross-cultural differences in mental health expressions on social media. *arXiv preprint arXiv:2402.11477*.

Hadeel Saadany, Constantin Orasan, Rocio Caro Quintana, Felix Do Carmo, and Leonardo Zilio. 2023. Analysing mistranslation of emotions in multilingual tweets by online mt tools. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 275–284.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 767–777.

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. Translating across cultures: Llms for intralingual cultural adaptation. *arXiv preprint arXiv:2406.14504*.

Emma Steigerwald, Valeria Ramírez-Castañeda, Débora YC Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D Tarvin. 2022. Overcoming language barriers in academia: Machine translation tools and a vision for a multilingual future. *BioScience*, 72(10):988–998.

Zhen Tan, Dawei Li, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *ArXiv*, abs/2402.13446.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. *https://aclanthology.org/2020.coling-main.384*.

Cynthia Van Hee. 2017. *Can machines sense irony?: exploring automatic irony detection on social media*. Ph.D. thesis, Ghent University.

Anning Wang, Qiang Zhang, Shuangyao Zhao, Xiaonong Lu, and Zhanglin Peng. 2020. A review-driven customer preference measurement model for product improvement: sentiment-based importance–performance analysis. *Information Systems and E-Business Management*, 18(1):61–88.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Warren Weaver. 1952. Translation. In *Proceedings of the conference on mechanical translation*.

Elda Weizman. 2022. Explicitating irony in a cross-cultural perspective: Discursive practices in online op-eds in french and in hebrew. *Contrastive Pragmatics*, 4(3):437–465.

World Values Survey Association. 2023. Inglehart-welzel world cultural map. Based on World Values Survey Wave 7 2023.