Benchmarking Machine Translation with Cultural Awareness

Binwei Yao[1], **Ming Jiang**[2], **Tara Bobinac**[1], **Diyi Yang**[3], **Junjie Hu**[1]
[1]University of Wisconsin-Madison, [2]Indiana University Indianapolis, [3]Stanford University
binwei.yao@wisc.edu, mj200@iu.edu, bobinac@wisc.edu
diyiy@stanford.edu, junjie.hu@wisc.edu

Abstract

Translating culture-related content is vital for effective cross-cultural communication. However, many culture-specific items (CSIs) often lack viable translations across languages, making it challenging to collect high-quality, diverse parallel corpora with CSI annotations. This difficulty hinders the analysis of cultural awareness of machine translation (MT) systems, including traditional neural MT and the emerging MT paradigm using large language models (LLM). To address this gap, we introduce a novel parallel corpus, enriched with CSI annotations in 6 language pairs for investigating **C**ulturally-**A**ware **M**achine **T**ranslation—**CAMT**.[1] Furthermore, we design two evaluation metrics to assess CSI translations, focusing on their pragmatic translation quality. Our findings show the superior ability of LLMs over neural MTs in leveraging external cultural knowledge for translating CSIs, especially those lacking translations in the target culture.

1 Introduction

Machine translation (MT) systems have achieved remarkable success in recent years, thanks in part to the pre-trained backbones of multilingual language models (Aharoni et al., 2019) and the availability of multilingual corpora (NLLB Team et al., 2022). Despite these advances, terminology translation remains challenging in both general contexts (Dinu et al., 2019) and specific domains like medicine and law (Ghazvininejad et al., 2023). Many existing MT studies on terminology translation have focused on breaking language barriers rather than cultural barriers, often assuming that literal (i.e., word-for-word) translation pairs already exist for the common knowledge shared by speakers of both the source and target languages (Anastasopoulos et al., 2021). However, culture is deeply intrinsic to language, and language translation entails cross-cultural communication (Newmark, 1988; Fernández Guerra, 2012). Due to the diverse nature of knowledge and norms across cultures, many cultural-specific items (CSIs) related to food, clothing, art, and religion are rarely used by speakers outside the items' associated cultural group, with some items even not existing in certain cultures (Woolford, 1983; Persson, 2015). As a result, cultural-specific items usually lack available literal translations across languages, leading most MT systems to perform poorly on cultural-centered translations in real-world deployment (Akinade et al., 2023; Liebling et al., 2022). As shown in Figure 1, common errors, including copy and factual errors, are still made by the state-of-the-art MT systems (e.g., Google Translate and ChatGPT). More importantly, the nature of CSIs leads to difficulty in collecting high-quality, yet diverse parallel corpora at scale, hindering a systematic evaluation of both the traditional neural MT systems and the emerging MT paradigm using large language models (LLM).

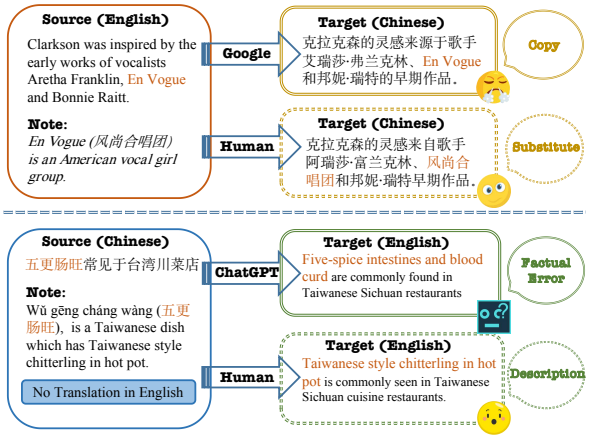

Figure 1: Culture-specific item translation errors.

To address this challenge, a handful of recent studies have begun to curate culture-related corpora for analysis from two main perspectives. The first perspective emphasizes regional varieties, such

[1]The corpus and code are released at https://github.com/BigBinnie/Benchmarking-LLM-based-Machine-Translation-on-Cultural-Awareness

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 13078–13096

as the variety of Portuguese used in Brazil versus Portugal (Riley et al., 2023). The second perspective focuses on cultural content in translation, such as recipes (Cao et al., 2024) and idioms (Li et al., 2024). Given the difficulty of obtaining word-for-word translation pairs for CSIs, these studies are shifting from the traditional *literal translation* paradigm to *free translation* which aims to convey the meaning of source texts and prioritizes readability and cultural relevance over strict accuracy and structural fidelity. Despite the valuable contributions of these works, two major concerns may still hinder the analysis of MT systems in navigating cultural nuances. First, the demand for high-quality data requires costly human annotations, restricting these studies from scaling up their curated data resources in terms of size, language pair diversity, and cultural domain coverage (see Table 1). Second, common MT evaluation metrics designed for literal translation lead to a lack of reliable assessments of free translation quality.

In this study, we address the two aforementioned concerns with a particular focus on translating culture-specific items. Specifically, we introduce an annotation-efficient data curation pipeline that can freely gather a diverse, large-scale, culture-centered parallel corpus while ensuring data quality. The resulting corpus, called **C**ulturally-**A**ware **M**achine **T**ranslation (**CAMT**), encompasses ***6*** *language pairs, covering* ***6,983*** *CSIs across* ***18*** *concept categories from* ***235*** *countries and regions*. To facilitate automatic assessment emphasizing CSI-centered free translation, we propose two evaluation metrics: CSI-Match and pragmatic translation assessment (PTA). The CSI-Match metric evaluates the translation accuracy of isolated CSIs, independent of their sentence context. In contrast, the PTA metric assesses the comprehensibility of CSI translations within sentence contexts, emphasizing their pragmatic effectiveness in communication with native speakers of the target language.

Leveraging our CAMT corpus and designed evaluation metrics, we conduct a systematic analysis to investigate the capability of state-of-the-art neural MT and LLM-based MT systems in translating cultural content. First, we examine their efficacy with two popular terminology translation strategies that utilize the CSI dictionary. Our findings indicate that the terminology translation strategies greatly enhance the CSI translation accuracy for both neural MT and LLM-based MT systems. However, LLMs exhibit superior capability in leveraging the external dictionary compared to NMTs, particularly for CSIs that lack well-known translations. Next, to further examine LLMs' capability for integrating external knowledge into translations, we explore prompting strategies that incorporate the CSI explanations in the prompts. Our results show that incorporating CSI explanations in the prompts notably improves the pragmatic translation quality, especially for CSIs without direct translations. In summary, our contributions are as follows:

- We curate a diverse parallel corpus in six language pairs with rich cultural-specific item annotations using a highly automatic pipeline.
- We introduce two new evaluation metrics (CSI-Match and PTA) to assess translation quality regarding cultural nuances, particularly for terms lacking established translations.
- We examine both LLM-based MT and NMT systems using our dataset and metrics. Our results indicate that LLMs can effectively incorporate external cultural knowledge, thereby improving the pragmatic translation quality of CSIs.

2 Related Works

Culturally-Aware Machine Translation: As languages and cultures are highly intertwined, there is a growing desire to empower cultural awareness of MT systems (Hershcovich et al., 2022; Riley et al., 2023). However, as cultural nuances are subtle, collecting culturally sensitive data (Akinade et al., 2023) remains costly and time-consuming. Therefore, current work on cultural-aware translations is limited to specific domains and language pairs (Cao et al., 2024; Li et al., 2024). It is also challenging to perform a human-centered evaluation of the cultural nuances (Liebling et al., 2022; Li et al., 2024). Existing studies have proposed strategies to evaluate cultural awareness of traditional MT systems by grounding images (Khani et al., 2021), adapting entities (Peskov et al., 2021) or targeting at dilates (Riley et al., 2022). Different from existing culturally relevant MTs, we focus on evaluating the cultural awareness of MT by translating culture-specific items, a relatively underexplored area.

MT with Terminology Previous studies on machine translation with terminology focused primarily on generic domains (Dinu et al., 2019), or popular ones (e.g., law, medicine) (Ghazvininejad et al.,

2023). However, translating culture-specific items carries its own set of unique challenges because literal translations of CSIs may not exist in the target culture, making translation adaption crucial for target language readers to understand these terms (Vinay and Darbelnet, 1995). The adaptation can create semantic asymmetry between the source words and their translations, which makes traditional translation evaluation metrics focused on semantic alignment insufficient for cross-cultural translation (Hershcovich et al., 2022).

External Knowledge for MT: There have been multiple threads of research efforts on integrating external knowledge such as bilingual translation lexicons for neural machine translation systems, including probability interpolation of lexicons (Arthur et al., 2016; Khandelwal et al., 2021), data augmentation by back-translations (Hu et al., 2019), decoding with a phrase memory (Wang et al., 2017), and pre-training with an entity-based denoising objective (Hu et al., 2022). Despite their effectiveness, these methods require further fine-tuning of the original MTs. As the parameters of LLMs (e.g., ChatGPT) may not be accessible, we focus on tuning-free methods for integrating external knowledge in this study (Dinu et al., 2019).

LLM-based MT: Large language models, such as GPT-3 (Brown et al., 2020), have proven effective in machine translation for various high-resource languages (Hendy et al., 2023; Jiao et al., 2023). In particular, a few recent studies have investigated the performance of LLM-based MT, including formality control of translation outputs (Garcia and Firat, 2022), in-context translation ability during pre-training (Shin et al., 2022), and multilingual translation (Scao et al., 2022; Zhu et al., 2023). Moreover, previous work indicates that LLMs can integrate external knowledge in the context into translation (Ghazvininejad et al., 2023; Li et al., 2024). However, the exploration of LLM-based MT on leveraging cultural knowledge to translate culture-specific items is still lacking.

3 CAMT Dataset

To minimize the need for human efforts while still obtaining diverse, high-quality CSI-centered translation pairs across multiple languages and cultures, we rely on Wikipedia to collect the data. The overall workflow of our data collection includes (1) building a wiki-centered cultural taxonomy (§3.1); (2) curating parallel sentences containing culturally-relevant entities (§3.2); and, (3) augmenting geo-metadata (§3.3). Figure 2 displays an overview of our data construction pipeline.

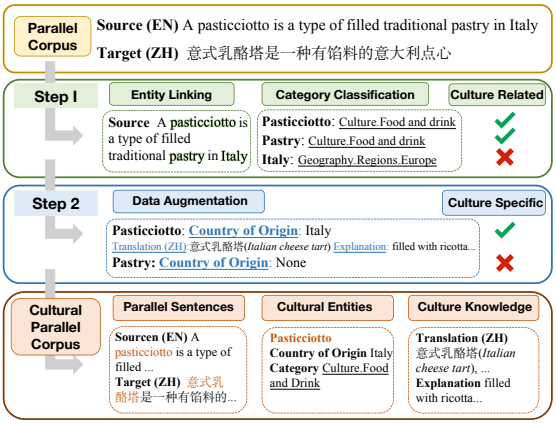

Figure 2: Overview of CAMT construction pipeline.

3.1 Cultural Taxonomy Extraction

Since culture is an abstract concept, it is difficult to capture fine-grained cultural characteristics from texts directly. With this consideration in mind, we referred to an existing CSI classification framework (Newmark, 1988), which has been popularly used in the study of human translations of cultural concepts, to identify culturally relevant texts from Wikipedia. Specifically, there are five CSI categories in this framework, including: 1) *ecology*; 2) *material culture*; 3) *social culture*; 4) *organizations, customs, ideas*; 5) *gestures and habits*. Each entity-centered Wikipedia page is labeled by a variety of Wikipedia categories (Asthana and Halfaker, 2018). To save the efforts of matching each entity on Wikipedia with each CSI category, we map CSI categories with Wikipedia categories by manually creating a mapping table (in Table 9) to establish connections between the two categories. Ultimately, 18 Wikipedia categories are identified as culturally related. An entity is classified as culturally related if the category of its Wikipedia page maps to one of the CSI categories.

3.2 Culture Parallel Text Collection

To construct a culture parallel text corpus (e.g., for English-Chinese), we collect public text articles from Wikipedia's translation tool that cover a wide range of cultural topics, and conduct sentence alignment to get parallel sentences (tools are detailed in Appendix §C). To expand the language coverage in our corpora, we also reuse open-source parallel corpora from OPUS (Tiedemann, 2016). These

Dataset	Language	Domains	Format
AppsCorpus (Kementchedjhieva et al., 2020)	en, es de, pl	Person Organization	Sent.
Adaption (Peskov and Hangya, 2021)	en, de	Celebrity	Ent.
IdiomKB (Li et al., 2024)	en, zh ja	Idiom	Sent.
CulturalRecipes (Cao et al., 2024)	en, zh	Recipes	Para.
CAMT (Ours)	en, zh, fr, es hi, ta, te	Cultural Categories	Sent.

Table 1: Dataset comparison. Sent., Ent., and Para. are abbreviations of sentence, entity and paragraph.

include Wikipedia v1.0 for English to French and Spanish, as well as Samanantar v0.2 for English to Hindi, Tamil, and Telugu. To identify sentences that contain culture-related items (Step 1 in Figure 2), we first perform entity-linking (Ringgaard et al., 2017) to identify Wikipedia entities on the source texts, then use classification tool (Asthana and Halfaker, 2018) to classify the Wikipedia categories of these entities. The categories are further mapped to our CSI categories using the cultural taxonomy (§3.1). The mapping table is shown in Appendix B. Finally, we only keep sentences that contain entities belonging to CSI categories.

3.3 Cultural Knowledge Augmentation

Existing MT studies (Arthur et al., 2016; Hu et al., 2022) have used external knowledge sources (e.g., Wikidata) to improve named entity translations. To enable future adaptations of these studies on our dataset, we parse Wikidata to augment cultural knowledge of CSIs (Step 2 in Figure 2), which includes their translations, descriptions, and aliases in multiple languages. We also obtain the plain text of the first paragraph of the Wikipedia article as the CSI explanation. Moreover, to identify cultural items that are specific to a certain country, we collect information on *country of origin* from Wikidata for each item and remove sentences that contain only items without an associated country of origin. We refer to these groupings of data for each CSI as CSI dictionaries, an example of which is shown in the data example in Appendix §A. This meticulous approach enriches our dataset with supplementary cultural knowledge, enabling us to evaluate MTs' performances in handling CSIs.

3.4 Dataset Analysis

We briefly compare CAMT with existing datasets that similarly focus on translating culture-specific content in Table 1. Similar to AppsCorpus and IdiomKB, translation pairs in CAMT are at the sentence level, aiming to provide fine-level textual context to explore the translation quality of CSIs from both semantic and pragmatic perspectives. Regarding data diversity, CAMT significantly expands the coverage to 18 cultural categories compared to prior work that focus on a specific domain. With respect to languages, CAMT includes 7 languages, which is more than in existing datasets (≤ 4).

We further conduct detailed corpus statistics on CAMT. As shown in Table 2, our dataset contains $6,948$ parallel sentences over 6 language pairs, of which $3,029$ sentences have CSI translations and the rest are non-translation instances. The total number of unique CSIs (called CSIs Types) in CAMT is $6,983$. Among various cultural categories, we find that *organizations, customs, ideas* and *material culture* are the top 2 categories, and *social culture* and *ecology* are the bottom ones. A more detailed breakdown of statistics of CAMT can found in Appendix §D.

4 Cultural Awareness Evaluation

To better capture the cultural nuances in CSI translations, we devise two evaluation metrics: (1) CSI-Match, which evaluates the accuracy of CSIs with labels, and (2) PTA, which assesses the pragmatic translation quality of CSIs without labels.

CSI-Match: Existing evaluation metrics for terminology are efficient for evaluating the accuracy of translations and the fluency of outputs (Anastasopoulos et al., 2021). However, previous metrics such as Exact Match (EM) assume that terminology translations must be exact matches, while reasonable adaptations of CSIs are also acceptable (Vinay and Darbelnet, 1995). To address this, we introduce the CSI-Match metric as a modification to the EM evaluation metric. CSI-Match measures the accuracy of term translation using a more nuanced, fuzzy matching approach. It calculates the maximal partial similarity ratio (PSR) between the reference CSI translations $t_1, t_2, \ldots, t_n$ and the system out-

Pair	Sent.	CSIs Counts	CSIs Types	CSI Translations
En-Zh	778	794	601	730
En-Fr	2,073	2,213	2,213	1,130
En-Es	1,580	1,652	1,652	817
En-Hi	1,086	1,127	1,127	168
En-Ta	677	695	695	118
En-Te	754	695	695	66
Total	6,948	7,176	6,983	3,029

Table 2: Dataset Statistics on Six Language Pairs.

put sentence S. CSI-Match is determined by Eq. (1), resulting in a value from 0 to 100. A higher value indicates a stronger similarity of the predicted CSIs to a set of CSI translation references.

$$\text{CSI-Match} = \max_{t \in \{t_1, t_2, ..., t_n\}} \text{PSR}(t, S) \quad (1)$$

PSR measures the maximum similarity between string t and any substring in S.

$$\text{PSR}(t, S) = \max_{s \in P} (1 - d(t, s)) \times 100 \quad (2)$$

$$P = \{S_{i:j} \mid 0 \leq i \leq j < |S|\} \quad (3)$$

where $S_{i:j}$ is a substring of S from word index i to j, and $d(,)$ is the normalized Levenshtein distance which measures the minimum number of insertions, deletions, and substitutions required to change one string into another (Levenshtein et al., 1966).

Pragmatic Translation Assessment (PTA): Existing evaluation metrics like BLEU (Papineni et al., 2002a) and COMET (Rei et al., 2020a) mainly focus on surface-level or semantic-level accuracy between the source and target texts. However, in the context of CSI translation where translation quality is tightly associated with target culture (Hershcovich et al., 2022), pragmatic translation quality becomes crucial. For example, a free translation based on the description of the CSI might have better pragmatic translation quality than a direct literal translation, as it would be easier for people from the target culture to understand. Therefore, we design a new assessment metric called *PTA*, measuring the win rate at which CSI translations by the MT system are judged to exhibit better pragmatic translation quality compared to human reference translations. Specifically, in the human evaluation, we specify what the CSIs are in the source language and ask native speakers of the target language to compare the CSI translations within the sentential context between an MT system and a reference, and then select the translation that is easier to understand. To improve the applicability of PTA when native speakers are not available, we use GPT-4o to replace human judgments for PTA, which has proven effective for the automatic evaluation of generative models in recent studies (Rafailov et al., 2023; Kocmi and Federmann, 2023). The evaluation prompt, shown in Appendix §E, is used for both human annotators and GPT-4. A higher PTA score means the system translates the CSI in a more comprehensive manner than the reference sentence does, which might use other accurate but less understandable translations of CSIs.

5 Experimental Settings

To fully evaluate the efficacy of MT system translating CSIs, we compare NMT systems with LLM-based MTs (§5.1). Secondly, to investigate the performances of traditional terminology translation methods on CSI translations, we evaluate two dictionary-based terminology methods on open-sourced NMT and LLMs (§5.2). Moreover, to gauge the capacity of LLMs to leverage external knowledge, we evaluate 4 cultural knowledge prompting strategies on tuning-free LLMs (§5.3).

5.1 MT systems in Comparison

We evaluate the following MT systems:

- **NMTs**: We asses the NLLB 1.3B (NLLB Team et al., 2022) model, which is a state-of-the-art multilingual MT model. We also use the Google Translate engine in our comparison.
- **LLMs**: We examine ChatGPT (GPT-3.5-turbo-1106) and the open-source LLaMA2-7B for comparison, as both have been proven to be efficient multilingual MT tools (Zhu et al., 2023).

5.2 Dictionary-based Methods in Comparison

For the open-sourced models (i.e., LLaMA and NLLB) we experiment with two additional methods proven to be highly effective in terminology MT during inference (Dinu et al., 2019). Specifically, we employ 1) the **Append** method: append the CSI dictionary before the input, whose format is "<CSI_1>:<CSI_1 Translations>,...,<CSI_n>:<CSI_n Translations>[Source language]"; and 2) the **Replace** method: replace the CSIs in the source sentence with their translation in target language. For LLaMA2, we use the following prompt in 8 shots: [Source language]:[Source sentence] = [Target language]:[Target sentence], which is efficient for machine translation (Zhu et al., 2023).

5.3 Prompting Strategies in Comparison

We explore various prompting strategies to introduce cultural knowledge into LLM-based MT. Our strategies generate in-context examples to integrate additional cultural knowledge, which involves employing CSI dictionaries and CSI explanations. Table 3 shows examples of four prompting strategies.

- **Basic Instruction (BI)** The basic machine translation prompt of LLMs (e.g. ChatGPT).
- **External CSI Translation (CT)** To assess the impact of incorporating a CSI dictionary within

Strategy	Prompt
Basic Instruction (**BI**)	Translate the following English text to Chinese
CSI Translation (**CT**)	The Chinese translation of culture entities in the sentence is as following: *cannoli:*里考塔芝士卷*(Ricotta cheese rolls),* 奶油甜馅煎饼卷 *(Sweet Cream pancake rolls)* Translate the following English text to Chinese
CSI Explanation (**CE**)	The explanation of culture entities in the sentence is as following: *Cannoli are Italian pastries consisting of tube-shaped shells of fried pastry dough ...* Translate the following English text to Chinese
Self-Explanation (**SE**)	User: Please explain cannoli in [Source Sentence] LLM: [Explanation] User: According to your explanations to cannoli, only translate the following English text to Chinese
Source	They are also commonly available at Italian-American bakeries in the United States, alongside other Italian pastries like cannoli and sfogliatelle.
Knowledge	Translations:cannoli:里考塔芝士卷*(Ricotta cheese rolls)*, 奶油甜馅煎饼卷 *(Sweet Cream pancake rolls)* Explanation: Cannoli are Italian pastries consisting of tube-shaped shells of fried pastry dough ...

Table 3: Prompting strategy examples (**Top**) and a source with cultural knowledge for En-Zh translation (**Bottom**).

the prompts, we include CSIs along with their corresponding translations prior to the basic translation instruction BI.

- **External CSI Explanation (CE)** CSIs may not have a direct equivalence in the target language's culture. Therefore, it becomes necessary to translate based on the explanation of the CSI to assist the target audience in better understanding the content. To assess the impact of explanations, we include the CSI explanation obtained from Wikipedia in the prompt before the basic translation instructions.
- **Self-Explanation (SE)** We also examine LLMs' internal knowledge for explaining the meaning of CSIs. Inspired by Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022), we treat the explanation of CSIs in a source sentence as an intermediate reasoning step before translating the whole sentence. We design the explanation prompting strategy in two steps for machine translation. First, we prompt the LLM to explain the meaning of all CSIs in the source sentence. Second, we ask the LLM to translate the whole sentence by combining the LLM's explanation with another prompt instruction.

6 Results and Analysis

In this section, we 1) compare the CSI translation performances of LLM-based MT systems with that of NMT systems (§6.1); 2) evaluate dictionary-based terminology translation methods to explore if they work on CSI translations (§6.2); 3) compare four prompting strategies of LLM-based MTs to explore how different prompts affect the LLMs' cultural awareness (§6.3); 4) conduct human evaluation to verify the correlation between automatic evaluation metrics and human assessment (§6.4).

6.1 Evaluating LLM-based MT v.s. NMT

To compare the cultural awareness of LLM-based MT systems with that of NMT systems, We employ two automatic metrics (CSI-Match and PTA) to evaluate 4 MT systems: the vanilla NLLB and Google Translate, and the BI prompting of LLaMA2 and GPT-3.5.

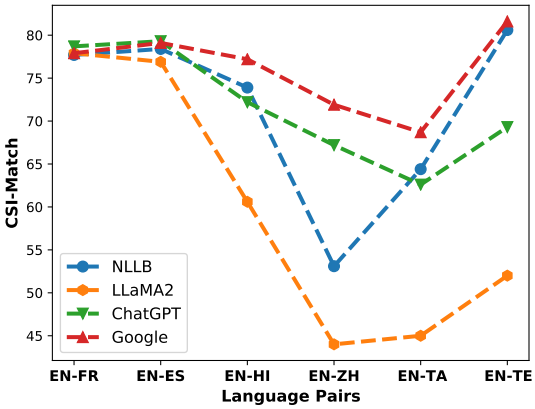

Figure 3: CSI-Match results on six language pairs.

NMTs Excel in CSI-Match on Low-resource Languages We evaluate the accuracy of CSI translations using CSI-Match across six language pairs, as shown in Figure 3. For the two Romance languages (French and Spanish), the performances of four MT systems are quite similar. However, Google Translate generally exhibits more consistent performance in non-Romance languages compared to LLM-based MT. NLLB's poor performance on EN-ZH is due to its limited translation capacity on EN-ZH, as validated in previous bench-

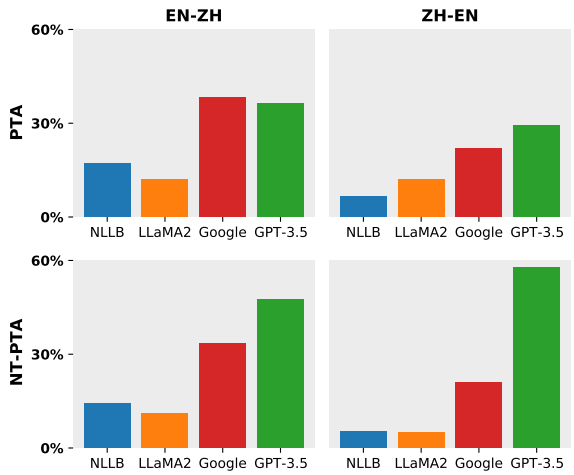

Figure 4: PTA results on English-Chinese translations.

marking work (Aharoni et al., 2019; Akinade et al., 2023). Specifically, NLLB and Google Translate outperform LLaMA2 and GPT-3.5 in translating three Indian languages: Hindi, Tamil, and Telugu. Additionally, for low-resource languages like Tamil and Telugu, the translation performance of LLM-MTs remains limited on traditional translation metrics (see Table 10). Therefore, LLM-MTs cannot yet be considered efficient multilingual translation tools for these low-resource languages.

GPT-3.5 Produces Better Pragmatic Translation on CSIs with No Established Translations. Given the cost of evaluation by commercial tools and human experts across all language pairs, we focus on English-Chinese in both translation directions when evaluating the pragmatic translation quality by PTA. Figure 4 presents the PTA assessed by GPT-4o for four MT systems' output compared to the reference sentences. In addition to PTA across the entire dataset, we separately evaluate PTA of samples containing CSIs with no established translations (**NT-PTA**). Notably, GPT-3.5 performs better than any other MT systems on NT-PTA, which potentially suggests that the instruction-tuning of LLMs beyond the translation task may be beneficial for the model to generate free translations that are easily understood by the target culture, especially for non-translation CSIs.

6.2 Evaluating Dictionary-based Methods

LLaMA is More Robust at Leveraging CSI Dictionaries than NLLB. We evaluate two dictionary-based terminology methods on NLLB and LLaMA for English-Chinese translations, as shown in Table 4. We find that straightforward

Model		Method	CSI-Match	PTA
NLLB	EN-ZH	**Vanilla**	53.1	17.1
		Append	58.3▲	17.7▲
		Replace	**78.7**▲	**20.5**▲
	ZH-EN	**Vanilla**	64.9	6.7
		Append	65.5▲	4.6▼
		Replace	**79.8**▲	**9.2**▲
LLAMA2	EN-ZH	**Vanilla**	45.0	12.1
		Append	**80.2**▲	**16.8**▲
		Replace	67.1▲	15.3▲
	ZH-EN	**Vanilla**	70.9	12.0
		Append	**85.6**▲	**17.5**▲
		Replace	80.6▲	16.5▲

Table 4: Evaluation of traditional dictionary-based methods on English-Chinese translations. ▲/▼ means the score is better or worse than the vanilla model.

strategies using dictionaries of CSIs, such as Replace and Append are effective for both NLLB and LLaMA on metrics that rely on string-matching (i.e. CSI-Match), as well as other semantic matching metrics (see Table 10). However, the appending strategy significantly benefits LLaMA more than NLLB. This suggests that LLaMA's ability for in-context learning and instruction-following enables the flexible integration of cultural knowledge at test time, a capability not present in traditional NMT systems like NLLB. Furthermore, traditional terminology translation methods can improve the PTA across the entire dataset. Without dictionaries, they still encounter challenges in improving the comprehensibility of translations that contain CSIs.

6.3 Evaluating Prompting Strategies

Model	Method	CSI-Match	PTA	NT-PTA
GPT-3.5	**BI**	66.2	33.2	31.7
	CT	**84.0**▲	35.6▲	-
	CE	67.1▲	35.8▲	**41.3**▲
	SE	67.7▲	**36.7**▲	36.5▲
LLaMA2	**BI**	43.7	11.3	11.1
	CT	**82.2**▲	**16.6**▲	-
	CE	43.3▼	10.8▼	**15.9**▲
	SE	47.8▲	10.3▼	12.7▲

Table 5: Evaluation of different prompting strategies on English-to-Chinese translations. ▲/▼ means the score is better or worse than the vanilla model.

LLM-based MTs open up the opportunity to incorporate free-form external knowledge to enhance the pragmatic translation quality of CSIs, especially for those without dictionaries. We ex-

Strategy	Outputs	PTA of GPT-4o	PTA of Human
BI	就像意大利的polenta concia一样，它可以作为主菜食用。	Lose	Lose
CT	和在意大利一样，波伦塔(*transliteration*) 在意大利被当作主菜。	Lose	Lose
CE	就像意大利的奶酪玉米粥(*cheese corn porridge*)一样，它可以作为主菜食用。	Win	Win
SE	就像意大利的奶酪玉米粥(*cheese corn porridge*)一样，它可以作为主菜食用。 Explanation by GPT-3.5: Polenta is a traditional Italian dish that originated in Northern Italy. It is a type of porridge made from cornmeal, and is similar in consistency to grits or cornmeal mush.	Win	Win
Source **Reference**	Just like polenta concia in Italy, it is eaten as a main dish. 就像意大利的玉米粥 (*corn porridge*)一样，它可以作为主菜食用。		
Knowledge	Translation: 波伦塔 (*transliteration*) Explanation: Polenta is a dish of boiled cornmeal that was historically made from other grains. The dish comes from Italy. It may be served served as a hot porridge.		

Table 6: The output example of four prompting strategies on GPT-3.5 for En-Zh translation.

plore various prompting strategies by integrating additional cultural knowledge, including dictionaries and explanations, to improve translations. We compare different prompting strategies for English-to-Chinese translations, employing 2-shot prompting approaches to obtain results from GPT-3.5 and LLaMA2. Table 5 shows the evaluation results.

External CSI Knowledge Improves LLM-MT. When comparing the strategies of using external knowledge in prompts (i.e., CT and CE), we observe that LLMs can effectively leverage both direct translations and indirect explanations. Specifically, CT enhances the CSI-Match score for CSI translations. However, CT is not effective when CSIs have no existing translations. In contrast, the 2-shot CE approach using GPT-3.5 improves NT-PTA from 31.7 to 41.3 in English-to-Chinese translations. This suggests that CSI explanations can notably aid in translating CSIs, particularly those without well-known translations. For LLaMA, the PTA score is similar to the baseline (with differences of fewer than 10 examples out of 778 data points) due to the limitations of LLaMA2-7B's capacity for English-to-Chinese translations. However, CE and SE approaches with LLaMA still show an improvement in NT-PTA, indicating that LLaMA2-7B can also leverage external explanations to improve the translation quality of CSIs without existing direct translations.

LLMs' CSI Explanations Also Help. We use SE to elicit LLMs' internal knowledge and find that the 2-shot SE approach with GPT-3.5 improves translation performance across all metrics for English-to-Chinese translations. This suggests that GPT-3.5 may already possess a significant amount of cultural knowledge about CSIs and can integrate this knowledge into the translation process. For LLaMA2, the PTA of SE is close to the baseline, and the improvement in NT-PTA is also limited, indicating that LLaMA2-7B may not have sufficient cultural knowledge of CSIs for English-to-Chinese translations.

Prompting with CSI Explanations Encourages LLMs to Do Free Translations. In Table 6, we provide examples of the CSI "polenta", an Italian corn porridge. Its Chinese translation on Wikidata is merely a transliteration. Under the CT strategy, GPT-3.5 directly copies this transliteration into the output, which may be considered correct but not comprehensible for native speakers of Chinese. In contrast, using the CE strategy, GPT-3.5 integrates the CSI explanation into the translation, freely translating the term "corn porridge" into Chinese. This makes it easier for readers to understand the nature of "polenta". Furthermore, The SE strategy successfully generates an explanation for "polenta" and incorporates it into the translation as "corn porridge", which leads to better comprehension for Chinese native speakers, as is reflected in the ratings of GPT-4o and the human annotator.

6.4 Human Evaluation

Metrics	Human Acc.	Human PTA
BLEU	79.6	86.2
BLEURT	80.6	86.6
COMET	77.8	89.3
Exact-Match	77.0	81.7
CSI-Match	**88.7**	90.0
GPT-4o PTA	87.1	**95.7**

Table 7: Pearson's Coefficients between automatic metrics and human evaluation on CSI translation

To compare the consistency between automatic metrics and human evaluation, we conduct a human evaluation on a subset of 200 English-to-Chinese translations. We assess the outputs from eight MT

systems: NLLB, LLaMA2, Google Translate, and GPT-3.5 across zero-shot BI, and two-shot BI, CT, CE, and SE settings. The native Chinese speaker evaluates the accuracy and PTA of the outputs. We then calculate Pearson's correlation coefficients between automatic evaluation metrics and human assessments. Specifically, we compare the performance of CSI-Match and PTA with four traditional automatic evaluation metrics: BLEU, BLEURT, COMET, and Exact-Match. The results, presented in Table 7, indicate that CSI-Match exhibits the highest correlation with human accuracy, while GPT-4o PTA shows the highest correlation with human PTA. These findings suggest that CSI-Match and PTA are effective evaluation metrics for assessing the translation quality of CSIs, which can better capture cultural nuances than traditional metrics.

7 Conclusion

To advance culturally-aware machine translation, we curate a high-quality, diverse parallel corpus (CAMT) with rich CSI annotations in 6 language pairs using an automated pipeline. We introduce two evaluation metrics, CSI-Match and PTA, to assess translation quality concerning cultural nuances. Our evaluation of LLM-based MT and NMT systems using CAMT reveals that LLMs can effectively incorporate external cultural knowledge, enhancing the pragmatic translation quality of CSIs. Our work provides essential data sources and insights for advancing culturally-aware machine translation, laying the groundwork for future investigation in this field.

Limitations

Language Pairs in Evaluation Our work takes a significant step toward toward understanding and evaluating the cultural awareness of machine translation on CSIs. We provide a culturally sensitive parallel corpus with rich annotations on cultural-specific items in six languages pairs. However, due to the cost of evaluation by commercial LLMs and human experts across all language pairs, we conduct parts of our experiments on English-Chinese translations, whose data quality is also verified by human experts. Building on our insights into English-Chinese translation, we hope to encourage future work to verify our findings on other language pairs, and we will release our code repository to streamline further investigations.

Cultural-Awareness Definition In this study, we focus on the evaluation of cultural-specific items (CSIs). However, evaluating cultural awareness beyond individual entities also deserves further investigation. Besides CSIs, many other types of cultural errors persist in the translation process, such as those related to linguistic style and slang (Hershcovich et al., 2022). Our work aims to mitigate cultural errors by starting with CSIs, promoting advancements in culturally-aware machine translation datasets, models, and evaluation methods. This is crucial for enabling machine translation to play a larger role in cross-cultural communications.

Evaluation by LLM Recent research has shown that GPT-4 demonstrates a high correlation with human experts in evaluating generation performance (Rafailov et al., 2023; Kocmi and Federmann, 2023; Li et al., 2024). However, using GPT-4 as an evaluator may still pose fairness issues due to internal biases and unbalanced language capabilities of LLMs. In this study, we aim to advance beyond traditional semantic alignment evaluation metrics to assess pragmatic translation quality in English-Chinese translations using GPT-4. Further investigation is needed to improve GPT-4's effectiveness as an translation evaluator.

Prompting strategies We only try 4 prompting strategies in our study, due to our work's focus on benchmarking the cultural awareness of current LLM-based MT systems. In the future, we'll test other methods, such as instruction tuning, to improve the performance of LLM-based MT.

Ethical Considerations

Although our study designs a suite of simple but effective prompting strategies to enhance the cultural awareness of LLM-based machine translation, we still observe the weakness of LLM-based translation on cultural concepts in certain regions (e.g., Asia) and hallucinations on low-frequency entities. Potential usage of these LLM translation outputs may still result in the spread of misinformation. Before deploying our methods to create reliable content such as creating translations of Wikipedia articles, practitioners should ensure another round of human post-editing. During the annotation process, the annotators (native speakers of the target languages) consist of the authors of this article, who know the goals of the study clearly.

Acknowledgements

We sincerely appreciate the valuable feedback provided by our reviewers, which greatly helped to improve the manuscript. BY and JH are supported by the Wisconsin Alumni Research Foundation. MJ is partially supported by the National Science Foundation (IIS-2438420). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Idris Akinade, Jesujoba Alabi, David Adelani, Clement Odoje, and Dietrich Klakow. 2023. Varepsilon kú mask: Integrating Yorùbá cultural greetings into machine translation. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 1–7, Dubrovnik, Croatia. Association for Computational Linguistics.

Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, Vassilina Nikoulina, et al. 2021. On the evaluation of machine translation for terminology consistency. *arXiv preprint arXiv:2106.11891*.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. *arXiv preprint arXiv:1906.01105*.

Ana Fernández Guerra. 2012. Translating culture: problems, strategies and practical realities. *Sic: časopis za književnost, kulturu i književno prevođenje*, 3(1):0–0.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *arXiv preprint arXiv:2202.11822*.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

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

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2022. DEEP: DEnoising entity pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1753–1766, Dublin, Ireland. Association for Computational Linguistics.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Yova Kementchedjhieva, Di Lu, and Joel Tetreault. 2020. The apposcorpus: A new multilingual, multi-domain dataset for factual appositive generation. *arXiv preprint arXiv:2011.03287*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Nikzad Khani, Isidora Tourni, Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. Cultural and geographical influences on image translatability of words across languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 198–209, Online. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. 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.

Daniel Liebling, Katherine Heller, Samantha Robertson, and Wesley Deng. 2022. Opportunities for human-centered evaluation of machine translation systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 229–240, Seattle, United States. Association for Computational Linguistics.

Peter Newmark. 1988. *A textbook of translation*, volume 66. Prentice hall New York.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ulrika Persson. 2015. Culture-specific items: Translation procedures for a text about australian and new zealand children's literature.

Denis Peskov and Viktor Hangya. 2021. Adapting entities across languages and cultures. *Findings of the Association for Computational Linguistics: EMNLP 2021*.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Parker Riley, Timothy Dozat, Jan A Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2022. Frmt: A benchmark for few-shot region-aware machine translation. *arXiv preprint arXiv:2210.00193*.

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.

Michael Ringgaard, Rahul Gupta, and Fernando CN Pereira. 2017. Sling: A framework for frame semantic parsing. *arXiv preprint arXiv:1710.07032*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186, Seattle, United States. Association for Computational Linguistics.

Jörg Tiedemann. 2016. Opus-parallel corpora for everyone. *Baltic Journal of Modern Computing*, 4(2).

Jean-Paul Vinay and Jean Darbelnet. 1995. *Comparative stylistics of French and English: A methodology for translation*, volume 11. John Benjamins Publishing.

Xing Wang, Zhaopeng Tu, Deyi Xiong, and Min Zhang. 2017. Translating phrases in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Ellen Woolford. 1983. Bilingual code-switching and syntactic theory. *Linguistic inquiry*, 14(3):520–536.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Data Examples

In Table 8, we present a data example from the English-Chinese corpus. Each data point consists of a pair of sentences. We meticulously annotate all culture-specific items (CSI) within the sentences. For each culture-specific item, we provide information including its category, country of origin, translations in the target language, descriptions in both the source and target languages, and an explanation. To illustrate the challenges that cultural-specific items pose for current Machine Translation (MT) systems, we provide translations from both Google Translate and ChatGPT for this example. It is noted that both Google and ChatGPT erroneously rendered the Chinese translation of "Wiener Schnitzel" as "pork chops" instead of the correct translation, which is "steak". This misinterpretation not only misleads Chinese readers but also introduces confusion to the entire sentence, whose meaning is "*The Shanghai-style pork chops are a twist on Austria's national dish, Wiener fried pork chops, which are more street food than steak.*"

Aspect	Content
Source (EN)	The Shanghai-style Fried Pork Chop is a modification from Wiener Schnitzel the national dish of Austria, and a fried pork chop is more a street food than a beef steak.
Target (ZH)	上海炸猪排的做法改良自奥地利国菜 维也纳炸牛排 *(Wiener fried steak)*，而炸猪排与牛排不同，它显得更加市井。
Cultural-Specific Item	Wiener Schnitzel
Category	Culture.Food and drink
Country of Origin	Austria
Translation (ZH)	维也纳炸牛排 *(Wiener fried steak)*
Description (EN)	breaded veal schnitzel
Description (ZH)	面包屑小牛肉炸肉排
Explanation	The entity, sometimes spelled Wienerschnitzel, is a type of schnitzel made of a thin, breaded, pan-fried veal cutlet. It is one of the best known specialities of Viennese cuisine, and one of the national dishes of Austria.
NLLB	上海风格的炸猪肉切片是从奥地利国家菜的维也纳施尼切尔*(transliteration)*改制而成,
LLaMA2	上海炒猪排是一种来自奥地利的牛肉炒肉块 *(Beef stir-fried cubes)*的改良型，而炒猪排更像是一道街头小吃而非牛肉炒肉块.
Google Translate	海派炸猪排是奥地利国菜 维也纳炸猪排 *(Wiener fried pork chops)* 的改良版，炸猪排与其说是牛排，不如说是街头小吃。
ChatGPT	上海式炸猪排是从奥地利的国菜 维也纳炸猪排 *(Wiener fried pork chops)* 改编而来，而炸猪排比牛排，更像是街边食物。

Table 8: A Data Example in the English-Chinese Corpus: in parentheses, we explain what the Chinese translation means.

CSI Category	Wikiproject Category	CSI Example
Material Culture	Culture.Food and drink	*Cotoletta (Italy)*
	Culture.Visual arts.Architecture	*the Summer Palace (China)*
	History and Society.Transportation	*Kan-Etsu Expressway (Japan)*
Social Culture	Culture.Sports	*RKC Waalwijk (Netherlands)*
	Culture.Media.Entertainment	*Far Rockaway (USA)*
Organisations, Customs and Ideas	History and Society.Politics and government	*Europe Ecology – The Greens (France)*
	Culture.Philosophy and Religion	*Fuller Theological Seminary (USA)*
	Culture.Literature	*Der Spiegel (German)*
	Culture.Visual arts.Visual arts*	*The Headless Horseman Pursuing Ichabod Crane (USA)*
	Culture.Visual arts.Fashion	*Bottega Veneta (Italy)*
	Culture.Visual arts.Comics and Anime	*Dragon Ball (Japan)*
	Culture.Performing arts	*Just Dance (USA)*
	Culture.Media.Music	*Trident Studios (UK)*
	Culture.Media.Films	*A Few Good Men (USA)*
	Culture.Media.Books	*Moby-Dick (USA)*
	History and Society.History	*Tusculum (Italy)*
Ecology	STEM.Biology	*Kapok (Netherlands)*
	Geography.Regions.*	*Qualicum Beach (Canada)*
Gestures and Habits	-	

Table 9: CSI vs. Wikiproject mapping table.

B CSI vs. Wikiproject Mapping Table

The mapping table between CSI definitions (5 categories in total) and Wikiproject categories[2] (18 categories) are shown in Table 9. Additionally, we provide examples for each category to clarify the respective meanings. The tool we used for Wikiproject category classification is drafttopic[3].

C Wikipedia Parallel Corpus Collection

To collect the English-Chinese parallel corpus from Wikipedia, we use the bilingual Wikipedia articles translated through Wikipedia's content translate tool[4]. This tool allows confirmed editors to translate Wikipedia articles from the source language to a target language with a machine translation system. By tracking their editing logs, we obtain the text triples consisting of the original text in a source language, the machine-translated text, and the human post-edited text in the target language. We then use a sentence alignment tool `bleu-align`[5] (Sennrich and Volk, 2010) to obtain a sentence-level parallel corpus. To obtain more language pairs, we reuse open-source data from OPUS, which includes Wikipediav1.0 [6] for English-French and English-Spanish, as well as Samanantarv0.2 [7] for English-Hindi, English-Tamil, and English-Telugu.

D Data Characteristics

Data Statistics Table 2 shows the statistics of our parallel corpora for the evaluation of MT systems on six language pairs. Particularly, for each language pair, we count the total number of detected CSIs by **CSIs Counts** and the number of unique CSIs by **CSIs Types**. It's noted that not all the CSIs have translations on Wikidata, so we determine the number of CSIs containing translations in WikiData by **CSI Translations**. Considering that many CSIs only exist within a specific culture group, which can't be located in the parallel corpus, CSIs that don't have a translation in other languages should take a higher proportion in the real-world corpus than in our dataset.

Data Diversity Culture is intricately linked to specific regions, and its manifestations can exhibit substantial variations across diverse regions and categories. Therefore, our dataset encompasses culturally specific items sourced from a wide array of regions and categories. Figure 5 shows the distribution of categories. Specifically, we mapped 18 Wikiproject categories into 5 culture categories. Since there is no Wikiproject category matching

[2] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Categories
[3] https://github.com/wikimedia/drafttopic
[4] https://en.wikipedia.org/wiki/Wikipedia:Content_translation_tool
[5] https://github.com/rsennrich/Bleualign
[6] https://opus.nlpl.eu/Wikipedia-v1.0.php
[7] https://opus.nlpl.eu/Samanantar-v0.2.php

the CSI category *gestures and habits*, we excluded this label from further consideration. Regarding the regions, we show the top 15 origin countries in our dataset in Figure 6. Among these regions, CSIs originating from English-speaking countries (e.g., the United States and the United Kingdom) have the highest representation. This is because we conduct entity-linking on the English source texts, resulting in a predominance of CSIs from English-speaking countries. However, the entity linking tool SLING[8] is multilingual, making it feasible to use our pipeline to include more CSIs from non-English speaking countries. This inclusive approach allows us to comprehensively evaluate the performance of machine translation models across a broad spectrum of cultural contexts.

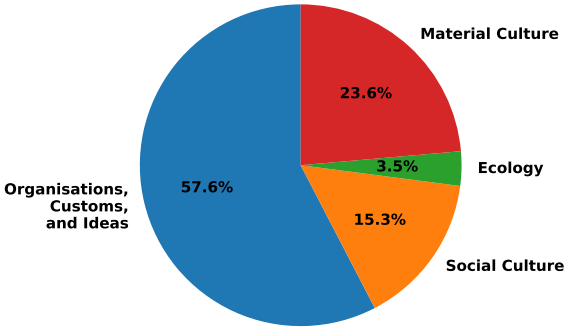

Figure 5: Category distribution on categories.

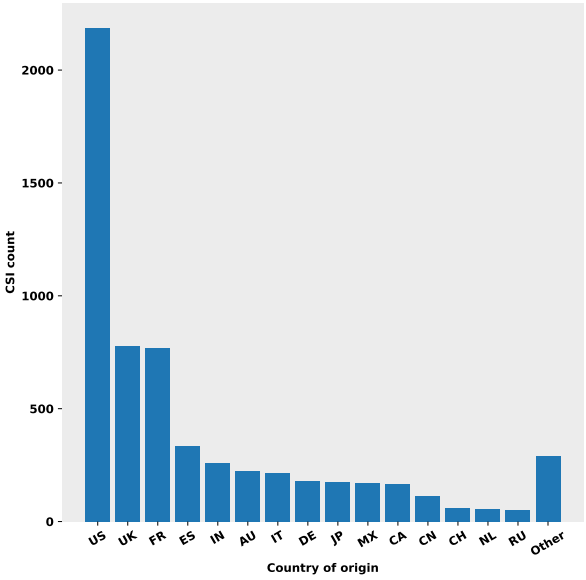

Figure 6: Data characteristics on regions.

[8] https://github.com/google/sling

E Evaluation Prompts of GPT-4o

It has been shown that GPT-4 can be an effective tool for evaluating the quality of generation tasks in DPO (Rafailov et al., 2023). We apply a similar prompt method for the pragmatic translation quality evaluation of CSIs. The prompt is as follows, with the system output and reference randomly shuffled into choices A and B:

Assuming you're a Chinese native speaker, which of the following translations has a more understandable translation in Chinese of following culture-specific item: "Goubuli"? Please only compare the items translation by ignoring the translation quality and length of the whole sentence.
<Source>
Translation A: <A>
Translation B: <B>
FIRST, provide a one-sentence comparison of the two translation, explaining which you prefer and why. SECOND, on a new line, if the translations of cultural-specific items: "Goubuli" in "A" and "B" are different, state "A" or "B" to indicate your choice, otherwise, use "C" to indicate your choice. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <"A" or "B" or "C">

For the human evaluation, we also use the same prompt as instructions to align the human evaluations with GPT-4o evaluations.

F Experiment Settings

The experiment settings of different models included in our paper are as follows:

- **NLLB** We use NLLB-200-1.3B-distilled[9] for our experiments. We use fairseq[10] to conduct the inference. The beam is set as 4, and the length penalty is set as 1.0.
- **LLaMA2** We use LLaMA-2-7B-hf[11] for testing. The sampling is set as True, leading to a multinomial sampling searching method.

[9] https://github.com/facebookresearch/fairseq/tree/nllb?tab=readme-ov-file
[10] https://github.com/facebookresearch/fairseq
[11] https://huggingface.co/meta-llama/Llama-2-7b-hf

- **GPT-3.5** We examine version gpt-3.5-turbo-1106. We use the ChatCompletion[12] API provided by OpenAPI For the generation, we set the parameters as default, for which the temperature is 1, top_p is 1, and frequency_penalty as 0.
- **GPT-4o** For GPT-4o, we use the latest version gpt-4o-2024-05-13 on Microsoft Azure platform by ChatCompletion, and we set the parameters as following: the temperature is 0 for a stable generation, top_p is 1, and frequency_penalty as 0.
- **Google translate** We call the Google Translate API[13] of Google Translate to get translations from it.

G Overall Automatic Evaluation

We evaluate the translation outputs using traditional automatic metrics such as BLEU (Papineni et al., 2002b), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020b). To be consistent with the evaluation method of NLLB, we calculate sp-BLEU (Goyal et al., 2022) for BLEU scores. In addition to traditional machine translation evaluation metrics, we also use CSI-Match to evaluate the translation quality of CSIs (described in §4). Table 10 shows the results of eight MT systems across six language pairs in two directions.

As shown in Table 10, both CSI dictionary incorporation (NLLB-A) and term replacement strategies (NLLB-R) enhance the translation quality of CSI for most language pairs, without significantly compromising the overall sentence translation regarding other metrics. Notably, NLLB-R outperforms other MT systems on CSI-Match, even including LLM-based MT. Interestingly, LLaMA2-7B shows an obvious drop in both traditional evaluation metrics and CSI-Match scores when translating English to three Indian languages and vice versa. One possible explanation is because of the insufficient Indian data during the pre-training of LLaMA2. Both CSI-involving translation strategies are beneficial for LLaMA-based translation. In non-Romance languages (i.e., Chinese, Hindi, Tamil, and Telugu), LLaMA2-A tends to yield better performances, whereas LLaMA2-R performs better in Romance languages (i.e., French and Spanish), which potentially suggests that injecting cultural knowledge through code-switching similar Romance languages works better than distant languages for LLM-based models. Furthermore, we assess the translation performances of ChatGPT and Google Translate. Both MT systems exhibit commendable performance in CSI translation, with Google Translate demonstrating superior translation results. Notably, Google Translate showcases consistent translation abilities, particularly in handling relatively low-resource languages like Tamil and Telugu.

[12] https://platform.openai.com/docs/guides/text-generation/chat-completions-api

[13] https://cloud.google.com/translate/docs/reference/rest

H PTA Evaluation Results Across Languages

We evaluate the PTA of two more language pairs. The evaluation result is shown in Figure 7. As with CSI-match on these two languages, the PTA performances of the 4 MT systems are pretty close. However, GPT-3.5 still shows superior performance on PTA compared to NMTs, indicating that GPT-3.5 has better capabilities to generate free translations for CSIs which can be easily understood by native speakers in the target culture.

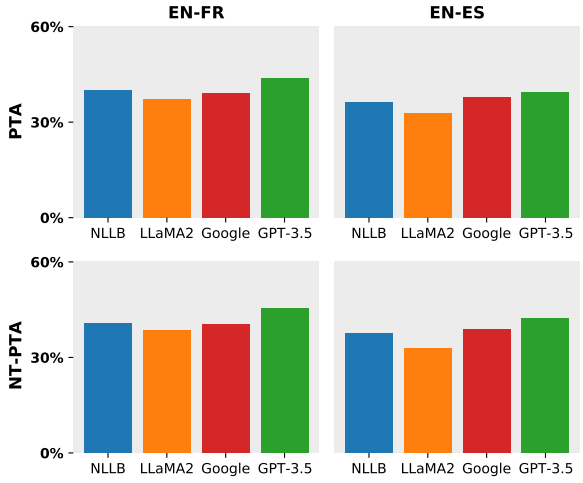

Figure 7: PTA results on En-Fr and En-Es translations.

I Generation Data Examples of Non-translation CSIs

Table 11 shows the results of the four prompting strategies on the CSI "milk toast" from English to Chinese, which has no known translations. Under the BI strategy, GPT-3.5 translates the American breakfast dish as "toast", failing to capture the defining feature of the dish, which is that it is soaked in milk. CT similarly fails, yielding the

Language Pair	Method	BLEU	BLEURT	COMET	CSI-Match	BLEU	BLEURT	COMET	CSI-Match
		EN-ZH				ZH-EN			
English-Chinese	**NLLB**	23.2	0.558	77.0	53.1	27.0	0.594	76.5	64.9
	NLLB-A	17.3	0.447	62.8	58.3	21.9	0.531	69.4	65.5
	NLLB-R	23.9	0.555	76.8	78.7	25.3	0.591	75.4	79.8
	LLaMA2	17.1	0.529	75.8	45.0	26.1	0.595	77.9	70.9
	LLaMA2-A	18.3	0.518	74.4	**80.2**	29.1	0.629	79.0	**85.6**
	LLaMA2-R	19.2	0.547	76.8	78.1	27.7	0.618	78.9	80.6
	ChatGPT	29.3	0.642	82.9	67.2	32.5	0.668	80.9	77.7
	Google	38.3	0.679	84.2	71.9	41.1	0.697	82.2	79.5
		EN-FR				FR-EN			
English-French	**NLLB**	37.4	0.585	78.3	77.7	36.3	0.634	77.8	88.9
	NLLB-A	37.4	0.582	78.0	77.4	35.4	0.628	77.2	88.3
	NLLB-R	37.0	0.577	77.8	**92.6**	36.6	0.635	77.7	92.1
	LLaMA2	34.4	0.558	76.2	77.9	27.6	0.462	66.0	72.5
	LLaMA2-A	35.0	0.568	67.8	82.7	34.7	0.633	77.7	92.8
	LLaMA2-R	34.7	0.550	75.6	89.8	30.2	0.620	76.8	**93.3**
	ChatGPT	36.2	0.594	78.9	78.7	37.6	0.629	77.8	89.8
	Google	31.1	0.573	77.5	77.9	36.1	0.632	77.2	88.9
		EN-ES				ES-EN			
English-Spanish	**NLLB**	48.8	0.707	83.4	78.4	50.7	0.718	83.5	90.3
	NLLB-A	48.3	0.705	83.3	78.7	49.9	0.713	83.0	90.8
	NLLB-R	47.9	0.696	82.9	**94.0**	50.9	0.717	83.4	**95.2**
	LLaMA2	44.6	0.679	81.6	76.9	45.3	0.704	82.6	89.9
	LLaMA2-A	44.5	0.674	81.2	82.0	46.6	0.706	82.7	93.3
	LLaMA2-R	44.5	0.675	81.3	93.0	44.2	0.702	82.3	94.7
	ChatGPT	47.9	0.711	83.6	79.3	50.7	0.712	83.4	92.3
	Google	42.9	0.704	82.0	79.1	49.8	0.722	83.0	90.9
		EN-HI				HI-EN			
English-Hindi	**NLLB**	32.8	0.637	0.747	73.9	38.4	0.683	83.5	93.7
	NLLB-A	32.1	0.630	0.734	78.7	38.4	0.676	82.8	90.9
	NLLB-R	32.9	0.639	0.748	**83.6**	38.4	0.684	83.5	**98.3**
	LLaMA2	7.3	0.438	0.493	60.6	12.6	0.441	63.8	67.5
	LLaMA2-A	7.0	0.439	0.493	81.9	18.9	0.546	74.3	79.1
	LLaMA2-R	8.2	0.444	0.502	80.0	14.8	0.480	67.7	67.7
	ChatGPT	24.1	0.592	0.701	72.2	32.0	0.649	81.4	93.4
	Google	33.8	0.651	0.753	77.2	39.9	0.690	83.0	94.5
		EN-TA				TA-EN			
English-Tamil	**NLLB**	25.8	0.706	83.0	64.4	31.5	0.645	80.4	92.1
	NLLB-A	25.5	0.698	82.3	70.5	31.0	0.639	79.8	94.9
	NLLB-R	25.8	0.707	82.9	81.6	31.5	0.645	80.4	**97.9**
	LLaMA2	1.7	0.309	39.6	44.0	4.3	0.331	51.9	54.4
	LLaMA2-A	1.2	0.341	39.6	**88.9**	9.1	0.417	61.5	87.0
	LLaMA2-R	1.4	0.321	40.0	73.9	5.1	0.305	53.7	68.7
	ChatGPT	10.4	0.496	62.5	62.6	16.9	0.539	72.2	87.4
	Google	26.9	0.712	82.7	68.7	31.4	0.649	80.4	91.9
		EN-TE				TE-EN			
English-Telugu	**NLLB**	31.4	0.628	81.1	80.6	34.7	0.643	81.2	87.3
	NLLB-A	29.2	0.624	80.4	81.9	34.5	0.635	80.5	89.8
	NLLB-R	31.4	0.628	81.1	**89.8**	34.7	0.643	81.2	**94.7**
	LLaMA2	3.2	0.207	41.0	52.0	0.9	0.190	41.6	33.2
	LLaMA2-A	4.0	0.244	42.3	88.8	5.8	0.356	56.7	78.0
	LLaMA2-R	4.0	0.237	42.0	77.0	2.3	0.269	48.8	44.1
	ChatGPT	16.8	0.484	67.3	69.3	23.3	0.567	74.8	78.8
	Google	32.7	0.635	81.0	81.6	34.9	0.653	81.2	89.5

Table 10: Automatic evaluation of all MT methods on six language pairs from both translation directions.

term “milk bread”. “Milk” as an adjective does not adequately describe the dish, and "bread" no longer specifies the toasted aspect. The former issue likewise arises with CE, a literal translation of “milk toast”. In contrast, using SE, GPT-3.5 integrates the CSI explanation into the translation, freely translating the term as “toasted bread soaked in milk”. This makes it easier for Chinese readers

to understand the meaning of "milk toast", as is reflected in the ratings of GPT-4 and the human annotator.

J Generation Examples of LLaMA

Table 12 shows the results of four prompting strategies on the CSI "burrito" from English to Chinese, defined as a *"flour tortilla wrapped into a sealed cylindrical shape around various ingredients."* Under the BI and CE strategies, LLaMA translates it as "bag" and "shell" respectively, failing to capture the essential feature of the dish, which is its rolled shape. The CT strategy successfully copies the dictionary translation. Interestingly, CE freely translates the word into "American southwest breakfast roll," accurately describing the food's shape. Additionally, CE prompts LLaMA to leverage related cultural knowledge to include the region description in the translation of the CSI.

K Performance Across Regions

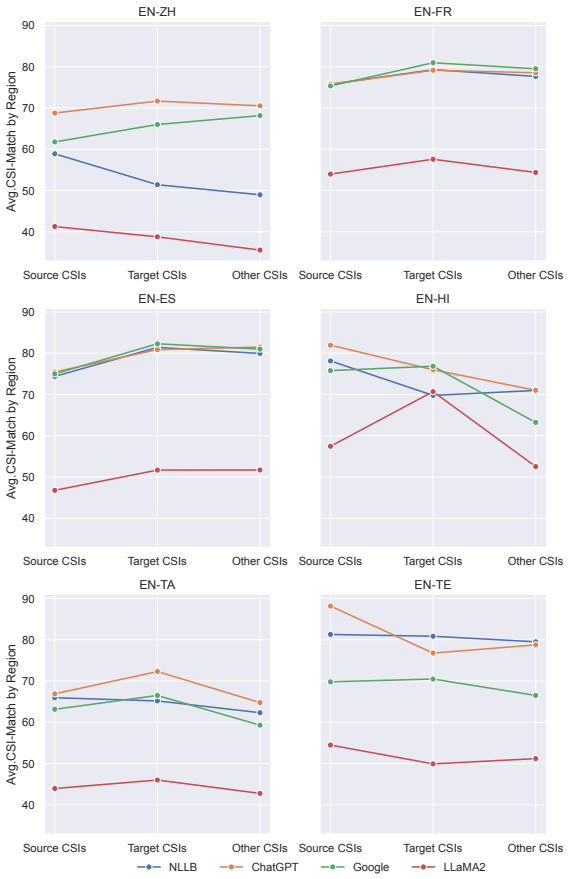

Figure 8: Avg.CSI-Match by regions.

Culture is often associated with a specific region, and its expressions can vary significantly across different regions and categories. To gain a deeper

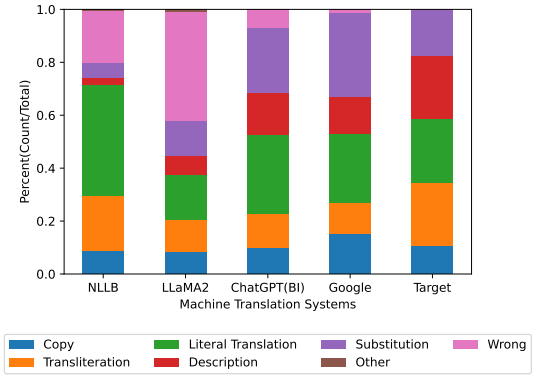

Figure 9: Percentage of translation strategies.

understanding of the influence of region on CSI translation, we categorized CSI into three groups: CSI originating from countries primarily using the source language, countries predominantly using the target language, and countries utilizing languages other than the source and target languages. In the six groups of English-to-XX translations, we calculated the average CSI-Match values of these three CSI groups respectively, shown in Figure 8.

Given that target CSIs must have the translation in the target language, translating target CSIs is akin to back translation. However, when translating the source CSI or other CSIs, the translation may either not exist in the target language or exist with lower word frequency. Consequently, the model is expected to yield better results for the target CSI. Surprisingly, our analysis reveals that most models excel at translating the target CSI back into the target language in Romance languages (i.e., French and Spanish). Notably, Google Translate consistently achieves superior translations across all languages. ChatGPT demonstrates better translation performance in Chinese and Tamil, while LLaMA2 succeeds in Hindi and Tamil for target CSI translation. In contrast, traditional translation models NLLB struggle with all non-Romance languages, failing to outperform the source CSI translation. This suggests that LLMs may possess enhanced learning capabilities for translating culture-related content. However, it is important to note that the current translation performance is not consistently stable.

L Comparison of Translation Strategies

Translation Strategies To explore potential factors benefiting pragmatic translation quality, we let a human annotator examine the models' translation

Strategy	Outputs	PTA of GPT-4o	PTA of Human
BI	该书中最终包含了1850种食谱，其中有烤面包(*toast*)。	Lose	Lose
CT	该书最终收录了1850个食谱，比如牛奶面包(*m*ilk bread)。	Lose	Lose
CE	这本书最终包含了1850个食谱，其中有牛奶土司(*l*iteral translation)。	Same	Same
SE	该书最终包含了1850个配方，其中有烤面包浸牛奶(*t*oasted bread soaked in milk)。	Win	Win
Source **Reference**	The book eventually contained 1,850 recipes including milk toast. 书中收录了1850个食谱，其中有牛奶土司 (*l*iteral Translation)。		
Knowledge	Translation: No existing Chinese translations Explanation: the entity is a breakfast dish consisting of toasted bread in warm milk, typically with sugar and butter...		

Table 11: Non-translation CSI output examples of prompting strategy on GPT-3.5 and a source-reference sentence pair with cultural knowledge for En-Zh translations.

Strategy	Outputs	PTA of GPT-4o	PTA of Human
BI	爱尔兰早餐卷的制作方式类似于一顿早餐休闲袋(*breakfast relaxing bag*)。	Lose	Lose
CT	爱尔兰早饭卷饼是与墨西哥卷饼(copy the dictionary)准确相似的。	Win	Win
CE	爱尔兰早餐卷的制作方式与美国西南部的早餐卷(American southwest breakfast roll)一样。	Win	Win
SE	爱尔兰的早餐卷是和早餐壳(*b*reakfast shell)。	Win	Win
Source **Reference**	The breakfast roll of Ireland is prepared similarly to a breakfast burrito. 爱尔兰的早餐面包卷制作方法亦类似于此 *(Replace the term with it)*。		
Knowledge	Translation:墨西哥卷饼 Explanation: The entity is a dish in Mexican and Tex-Mex cuisine, consisting of a flour tortilla wrapped into a sealed cylindrical shape around various ingredients...		

Table 12: CSI output examples of prompting strategy on LLaMA and a source-reference sentence pair with cultural knowledge for En-Zh translations.

strategies. We categorize the translation strategies of CSIs based on prior translation theories (Newmark, 1988; Persson, 2015). These theories define different categorizations of strategies to improve the comprehensibility of CSIs while maintaining cultural integrity. We select 4 strategies that are common to our dataset. They're 1) **Transliteration** that phonetically translates source CSIs; 2) **Literal translation** that directly translates word-by-word; 3) **Description** that integrates CSI descriptions of the CSIs into the translation; 4) **Substitution** that replace source CSIs by a semantically equivalent item in the target language; 5) **Copy** that directly copies the source language of CSI into the target language; 6) **Wrong** that indicates entirely incorrect translations; and 7) **Other** that employs other strategies in translation. Figure 9 shows the ratio of each strategy in four MT systems. We find that models with higher PTA (e.g., ChatGPT and Google translate) use description and substitution at a significantly higher rate, indicating that these two strategies help improve the understanding of CSI for target-language speakers. Notably, LLaMA2 incorporates a higher frequency of substitution and description methods compared to traditional NLLB. However, this increased diversity in translation output comes at the cost of reduced stability in the outputs. As a result, LLaMA2 tends to yield more inaccurate translations, whereas NLLB relies more on Literal Translation and Transliteration to translate CSIs.

Figure 10: Comparison of Translation Strategies: The value in each grid represents the win rate of the method on the x-axis in comparison to the method on the y-axis.

In order to further compare the impact of different translation strategies on pragmatic translation quality, we analyzed the comparison between

different translation strategies based on the ranking results of human evaluation. Specifically, we also rank the different translation strategies used by the MT systems according to the rank of the MT system's comprehensibility given by humans, which is shown in Figure 10. It's shown that the win rate of descriptions for all methods surpasses 0.5, and the win rate of substitutions, excluding descriptions, significantly exceeds 0.5. This implies that translations employing these two strategies are generally deemed more comprehensible by human annotators. Moreover, Literal Translation outperforms Transliteration, highlighting that transliteration may diminish the clarity of CSI in translation compared to a literal approach. Notably, the win rate of copying for both Literal Translation and Transliteration hovers around 50%, indicating that these two methods may introduce confusion, and their readability underperforms directly copying the original word.