Cultural Adaptation of Menus: A Fine-Grained Approach

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, Alexandra Birch
University of Edinburgh
zhonghe.zhang@hotmail.com, claire.xiaoyu.he@gmail.com
{vivek.iyer, a.birch}@ed.ac.uk

Abstract

Machine Translation of Culture-Specific Items (CSIs) poses significant challenges. Recent work on CSI translation has shown some success using Large Language Models (LLMs) to adapt to different languages and cultures; however, a deeper analysis is needed to examine the benefits and pitfalls of each method. In this paper, we introduce the ChineseMenuCSI dataset, the largest for Chinese-English menu corpora, annotated with CSI vs Non-CSI labels and a fine-grained test set. We define three levels of CSI figurativeness for a more nuanced analysis and develop a novel methodology for automatic CSI identification, which outperforms GPT-based prompts in most categories. Importantly, we are the first to integrate human translation theories into LLM-driven translation processes, significantly improving translation accuracy, with COMET scores increasing by up to 7 points. The code and dataset are available at https://github.com/Henry8772/ChineseMenuCSI.

1 Introduction

Translating restaurant menus is a challenging, non-literal translation task. Unlike other texts, dish names are not merely lists of ingredients and culinary methods; they are short (Pellatt and Liu, 2010) and culturally rich expressions that require an understanding of cultural traditions (Amenador and Wang, 2022), symbolism (Lam et al., 2018), and local nuances. This complexity is compounded by LLMs and Neural Machine Translation (NMT) systems that often lack the cultural awareness necessary to accurately understand these nuances (Liu et al., 2024; Naous et al., 2023; Tao et al., 2024). This results in mistranslations that can confuse and mislead the target audience (Garcea et al., 2023; Gallo et al., 2021), such as in Figure 1.

A key challenge in menu translation lies in the handling of Culture-Specific Items (CSIs), defined as "concepts that are specific to a particular language or group" (Aixelá, 1996). For example, the literal translation of the Chinese dish 蚂蚁上树 is "Ants Climbing a Tree" – but this is actually a figurative Chinese expression that should be translated in English as "Sauteed Vermicelli with Minced Pork". The Chinese name creatively expresses the idea that pork resembles ants, while vermicelli represents tree branches. Existing machine translation systems, trained on plain, sentence-level translations, fail to capture these cultural subtleties and generate literal translations (Figure 1).

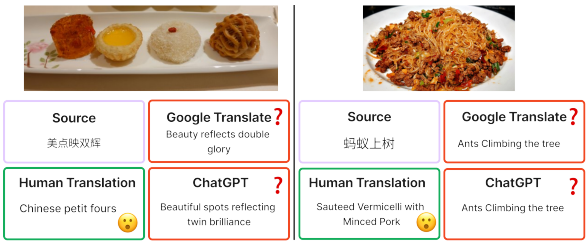

Figure 1: CSI translation errors by Google Translate and ChatGPT 3.5 in translating Chinese culinary terms.

However, there has been little work in NLP exploring CSI translations in-depth, particularly focusing on how the translation outputs generated by neural models should be improved. There has been foundational work on improving translations of CSIs by LLMs through enhanced prompting strategies (Yao et al., 2024). Simultaneously, there has also been work on adapting CSIs (Peskov et al., 2021; Cao et al., 2024; Singh et al., 2024), but their focus has been on adapting culture-specific named entities. In this work, we seek to go beyond entities, and approach the translation of figurative language imbued with cultural nuance, as exemplified in Figure 1 – which is quite underexplored.

In linguistics, cultural translation theories have been developed and widely adopted by human theorists and translators over decades. We aim to improve MT of CSIs by bringing the wisdom of Translation Theory research to modern NLP mod-

Proceedings of the Ninth Conference on Machine Translation, pages 1258–1271
November 15-16, 2024

els like LLMs. Our approach improves the identification and translation of figurative and culturally nuanced CSIs. Unlike previous CSI identification methods, our method does not depend on parallel corpora or extensive knowledge graphs (Yao et al., 2024; Han et al., 2023) – but at the same time, we also show how recipes can be *optionally* leveraged as a source of external knowledge to enhance performance even further. We also propose a novel CSI taxonomy for Chinese-English, that allows for a detailed analysis of figurative and culturally nuanced language and the translation challenges therein. We evaluate our proposed methods using a large dataset of Chinese dish names, ensuring robust and reliable results.

Our key contributions are as follows:

1. We introduce *ChineseMenuCSI*, a fine-grained dataset of 4,275 bilingual Chinese-English restaurant menu entries from UK Chinese restaurants. The dataset is categorised into CSI and Non-CSI entities, with 480 entries further bifurcated into specific CSI categories, enabling an in-depth analysis of CSI translation efficacy of LLMs and NMT systems.

2. We propose novel techniques for identifying CSIs, grounded in human translation theory. These techniques match or outperform current GPT-based prompts – all without needing external knowledge graphs or parallel corpora.

3. Lastly, we show how external knowledge, in the form of recipes, can add to the benefits of translation strategies and enhance CSI translation performance further - achieving significant improvements in COMET scores, with gains of +3 to +7 points across CSI categories.

The code and datasets are available at https://github.com/Henry8772/ChineseMenuCSI.

2 Related Work

2.1 CSIs in Translation Studies

Aixelá (1996) was among the first scholars to introduce the term "culture-specific items" (CSIs) to refer to elements in texts that are unique and significant in a specific culture. CSIs may include objects, classification systems, or measurement tools common in the source culture but foreign to the target culture. Additionally, CSIs can encompass transcriptions of opinions/habits specific to a culture, which are often reflected in the language structure, style, and content.

Culture is closely related to understanding and translating CSIs, as Aixelá (1996) highlighted. In the 1960s, Nida and Taber (2003) introduced the concepts of formal and dynamic equivalence in translation to distinguish between structurally accurate and fluency-focused approaches to translation. These concepts have laid the foundation for subsequent translation theories, including those related to cultural translation. Expanding on this, Newmark (1988) proposed a set of robust strategies for translating cultural elements, which have been particularly influential in translating Chinese culinary CSIs, as noted by Amenador and Wang (2022).

According to Newmark (1988), adaptation uses a recognized equivalent between two cultures. This strategy has been explored by Pellatt and Liu (2010) on Chinese menu translation. Newmark (1988) proposed three equivalent strategies for translation: cultural, functional, and descriptive – which we introduce later to improve LLM translation performance in §5.2.

Neutralisation is another translation strategy related to CSI translation. As proposed by Chou et al. (2016), on the continuum between foreignisation (focusing on source culture) and domestication (focusing on target culture), there are intermediary approaches, including neutrality and neutralisation. For culture-specific text, neutralisation involves paraphrasing to convey the meaning of a CSI. After analysing the translations of Chinese dish names into English, Amenador and Wang (2022) found that neutralisation, by substituting the source text element with a more or less detailed explanation of its meaning, is the most commonly used translation strategy by human translators for translating Chinese dish names into English.

In this paper, we use these conventional translation strategies employed by human translators as instructions in zero-shot prompts, to enhance CSI translation quality of LLMs (in §5.2).

2.2 Culture-Aware NMT

Despite the early successes of NMT (Bahdanau et al., 2015; Sutskever et al., 2014), translation of culture-specific texts has remained a daunting task. In addition to the challenge of translating rarer words and adapting to under-resourced domains (Koehn and Knowles, 2017), CSIs are deeply intertwined with cultures (Hershcovich et al., 2022; Liebling et al., 2022; Yao et al., 2024) – something

even the most capable neural models of today fail to grasp, particularly for non-Western cultures (Masoud et al., 2023; AlKhamissi et al., 2024; Nayak et al., 2024).

While there have been related works on domain-specific translation, including terminology translation (Dinu et al., 2019), disambiguation (Iyer et al., 2023a,b) and named entity translation (Hu et al., 2022), CSIs often lack direct equivalents in other languages, making translation complex and hard to understand cross-culturally (Yao et al., 2024).

Our approach uniquely combines translation studies with modern NLP techniques to identify and translate CSIs, resulting in more culturally sensitive and comprehensible translations.

2.3 Cultural Awareness and Adaptation in Large Language Models

In recent times, many works have shown that LLMs contain significant cultural biases against non-Western cultures (Cao et al., 2023; Liu et al., 2024; Masoud et al., 2023; Naous et al., 2023; Tao et al., 2024). In response, there has been a growing focus on improving cultural awareness in LLMs through prompt-engineering techniques (Wang et al., 2024; Tao et al., 2024) and fine-tuning on culture-specific data (Chan et al., 2023; Li et al., 2024a,b). Various tasks have been used to assess LLMs' cultural awareness, including tasks like culturally aware inference (Huang and Yang, 2023; Yao et al., 2024) and common sense reasoning on specific languages (Koto et al., 2024a,b).

Previous works on cultural awareness have primarily focused on understanding cultural norms in different languages rather than accurately translating culture-specific text. While well-explored in translation studies, cultural adaptation is rather understudied in NLP. Initial efforts in this direction have included adaptation of recipes (Cao et al., 2024) and localisation of named entities through adaptation (Peskov et al., 2021) or explicitation (Kementchedjhieva et al., 2020; Garcea et al., 2023; Han et al., 2023). Most similar to our work is that of Yao et al. (2024), who also released a CSI dataset covering 6 languages, on which they benchmark LLMs and NMT systems.

In contrast, our goal is to conduct a more fine-grained evaluation, given multiple CSI types in any given language. So, we leverage translation studies to create a dataset that classifies Chinese-English dishes into fine-grained categories, which we use for downstream evaluation, analysis and a detailed ablation of our proposed techniques. While we focus on the Chinese-English pair and culinary domain in this work, our framework and proposed techniques are agnostic of language/domain, and are designed to be easily scalable.

3 ChineseMenuCSI Dataset

We introduce a new bilingual Chinese-English Restaurant Menu (ChineseMenuCSI) dataset consisting of 4,275 human-verified dish entries collected from restaurants in UK.

3.1 Data Collection

We develop a Selenium-based web crawler[1] to gather Chinese menu translations from UK restaurant websites. After manually reviewing 50 restaurants, we selected those with ratings above 3 out of 5 and average meal prices over £20, ensuring higher-quality menus not generated by commercial Machine Translation systems like Google Translate. These restaurants were sourced from TripAdvisor[2].

Additionally, we developed a heuristic menu parser capable of accurately extracting structural content from image-based menus. Details of this parser are provided in Appendix A.2.

3.2 CSI Taxonomy

Translating Chinese menu items into English presents unique challenges because the dish names contain non-descriptive, picturesque elements (Pellatt and Liu, 2010). Our initial data inspection revealed that CSIs within these dish names contribute varying degrees of complexity to the translation process, and carry differing levels of figurativeness brought by cultural and linguistic nuances. Inspired by translation theory literature that tends to categorise Chinese dishes into concrete and abstract categories (Lam et al., 2018), we develop an approach to categorise the Chinese dish names in our dataset into three groups based on the degree of figurativeness in each CSI.

Category 1: Concrete CSIs (With a Low-level/No Figurative Meaning)

Definition: The CSIs in this category have a minimal figurative meaning, often referring to tangible attributes like ingredients, colour, taste, container, processing method, and dish appearance. Readers

[1]Selenium: https://www.selenium.dev/

[2]TripAdvisor: https://www.tripadvisor.com

can easily understand these dish names as the information is either shared between the source and target cultures or has widely used translations in the target culture.

Example: An example from the corpus is "咕噜猪肉" *(sweet and sour pork)*. The first two Chinese characters "咕噜" denote the Guangdong-style "sweet and sour" method, a culinary translation widely recognised outside the Chinese culture. The last two characters "猪肉" mean "pork", a culturally universal ingredient.

Category 2: Creative CSIs (With Some Figurative Meaning)

Definition: This category features dish names that blend concrete lexical terms with figurative meanings, creating inventive expressions that extend beyond literal definitions. Understanding these dishes necessitates integration of creative flair with concrete information, presenting challenges.

Example: "水煮鱼" (*Poached fish fillet with chilli oil and herb* or *Sichuan-style boiled fish*) originates from Sichuan, China. While the literal translation of the characters is "water-boiled fish", "water-boil" carries a creative description, representing the cooking state. This dish involves a Sichuan cooking style that uses hot chilli oil and Chinese herbs. "Poached fish fillet with chilli oil and herb" effectively describes the ingredients and cooking method, while "Sichuan-style boiled fish" adds cultural context by highlighting the dish's regional origin. Both are valid translations but different strategies are used.

Category 3: Abstract CSIs (With a High-level of Figurative Meaning)

Definition: This category encompasses dish names that exist beyond the realm of literal translation, and require in-depth cultural knowledge to understand. Crafted from metaphors, idioms, allegories, and other figurative language, these names disconnect from straightforward translations to engage in storytelling, aiming to convey broader narratives, evoke emotions, or reflect cultural heritage.

Example: "佛跳墙" *(Buddha Jumps Over the Wall* or *Steamed Abalone with Fish Maw in Chicken Broth)* metaphorically describes a dish so enticing that even a vegetarian and divine figure like Buddha would leap over a wall to taste it. Popular translations include "Buddha jumps over the wall" as a direct translation, and "Steamed Abalone with Fish Maw in Chicken Broth" includes ingredients and cooking methods, reflecting the dish's cultural and culinary nuances.

3.3 Data Annotation

To annotate our data, we first seek to classify the data into CSI and non-CSI entities, and if it is a CSI, we want to categorise it into one of the above-listed groups. Given the dataset has as many as 4.3K CSIs, we approach the annotation process in two stages: a) in Stage 1, we conduct a broad, albeit rough, annotation of the entire dataset by two volunteer annotators, and b) we uniformly sample from the annotations of Stage 1 to ensure a fair distribution across categories, and conduct a more focused and rigorous annotation process using five volunteer translators – to create our fine-grained test set. We describe these stages in detail below:

Stage 1 (Broad) Annotation: Two annotators, who are postgraduate students and professional Chinese-English translators, reviewed and labelled all 4,275 entries. Both are native Chinese speakers proficient in English, ensuring high linguistic and cultural expertise. Firstly, we classify the entries into CSIs and non-CSIs. We use Cohen's kappa (Cohen, 1960) to measure agreement between annotators and obtained a high score of 0.91 - likely because the classification of CSIs and non-CSIs is mostly unambiguous. For the 187 entries without consensus, we invited a third annotator to label and assigned the final label using a majority vote.

For entries with CSI, the annotators further categorised the dish into one of the three CSI categories. The annotation results are reported in Table 1.

Label	Category	Count
0	Non-CSIs	2003
1	Concrete CSIs	1658
2	Creative CSIs	494
3	Abstract CSIs	120

Table 1: Distribution of menu items across CSI taxonomy in the ChineseMenuCSI dataset.

Stage 2 (Focused) Annotation: In Table 1, we note that Category 3 is the smallest, with only 120 items. To evenly balance our test set, we randomly sample 120 items from each category: 0 (Non-CSIs), 1 (Concrete CSIs), 2 (Creative CSIs), and 3 (Abstract CSIs), totalling 480 items. This subset was annotated by a larger and more diverse group of five annotators, who, like the first-stage

annotators, included professional Chinese-English translators and postgraduate students – all native Chinese speakers proficient in English.

For span annotation, we first segment the dish name into spans and phrases using Jieba[3]. Annotators then label the spans that correspond to CSI. To assess inter-annotator agreement, we use Fleiss' kappa (Fleiss, 1971) across two levels: CSI fine-grained categorisation and span-level CSI identification; the results are summarised in Table 2.

Annotation	Kappa	Interpretation
CSI vs. Non-CSI	0.91	High
CSI Category	0.63	Substantial
CSI Identification	0.70	Substantial

Table 2: Inter-annotator agreement scores for different annotation tasks.

For the fine-grained CSI categorisation, we exclude items that do not attain majority consensus (at least 3 out of 5 annotators in agreement), resulting in a kappa score of 0.63. This score falls within the range of substantial agreement (0.6-0.8) but is lower than the CSI vs. Non-CSI score due to the subjective nature of the fine-grained categorisation. This level of agreement is comparable to ranges reported in related work (Huang and Yang, 2023; Soderstrom et al., 2021). Lastly, for CSI identification at the span-level, i.e. within a given dish name, the kappa score is 0.70 - which indicates substantial agreement as well.

4 CSI Automatic Identification

To accurately translate CSIs, it is essential to first identify which parts of the text comprise CSIs. Previous studies have approached this challenge in different ways. Han et al. (2023) focus on implicit detection and use a relative distance of terms in Wikidata, but it does not include all Chinese dish CSI. Yao et al. (2024) rely on parallel corpora with entity-linking to find CSIs; however, the approach is infeasible for online MT, where we need to identify CSIs beforehand to produce translations from the monolingual source text.

Inspired by these methods, we propose a method called **Combined CSI Identification**, that uses a combination of three checking criteria for CSI identification, and classifies CSI if at least two of the following three checks are met. The checks are Round-trip Translation (RTT), Cultural Uniqueness (CU), and Historical Significance (HS).

[3]Jieba: https://github.com/fxsjy/jieba

4.1 Round-trip Translation (RTT)

Since CSIs are defined as terms unique to a specific language or culture (Álvarez and Vidal, 1996), based on the assumption that they do not have corresponding translations in the target language, we propose using round-trip translation (RTT) as one of the identifying criteria.

1. **Initial Translation**: Translate the Chinese dish name to English using Google Translate.

2. **RTT Translation**: Translate the English version back to Chinese using DeepL Translate and split it into Jieba-segmented words. Using different translation systems for RTT proved most effective in identifying CSIs.

3. **Identification**: Subtract the segmented words in the RTT from those in the original text. The remaining words are potential CSIs.

$$\text{CSIs} = \text{Original Words} - \text{RTT Words}$$

   Using Jieba's cut-for-search module, which returns all words and phrases, a phrase is considered CSI only if all of its words are omitted in the RTT, and not otherwise.

This method has its limitations, for example, it could also: a) return words that are not CSIs and are just difficult to translate, and b) miss CSIs that have literal translations. We find in §6.3 that it performs strongly in identifying CSIs in most cases.

4.2 Cultural Uniqueness (CU)

According to Newmark (1988), "unfindable" words are often less frequently seen within a language. Words with cultural and historical references can be deeply embedded in a specific culture or history, making them rare or unfamiliar to outsiders.

We use Jieba to segment words in the ChineseMenuCSI dataset, then measure each word's frequency and calculate its inverse frequency. A cut-off at the 95th percentile of these inverse frequencies is set based on a manual review of 100 words. Words above this cut-off are marked as potential CSI. Words not previously seen are given an inverse frequency of 1, indicating they are potential CSI. No smoothing techniques are applied, as inverse frequency is used against a fixed threshold rather than for probability calculations.

4.3 Historical Significance (HS)

Chinese food names and CSIs often contain historical narratives such as historical events, figures and periods (Lam et al., 2018; Amenador and Wang, 2022). To identify these, we use the Wikipedia API to search for individual words or entire dish names. If a word's Wikipedia page includes a "History" section, it is considered a potential CSI. The words appearing 30 times or more, such as "chicken" or "sauce" are excluded as generic terms.

5 CSI Translation

Having identified CSIs, we propose prompting strategies to improve CSI translations. Our strategies fall into two categories: Recipe-based Translation and Translation Studies-inspired Prompting.

5.1 Recipe-based Translation

We explore using recipe information to improve the translation of CSI dish names. By incorporating the most relevant recipe as external knowledge, we experimented with two zero-shot prompt strategies: **Default Recipe** prompting and **Recipe + Explain-then-Translation** prompting (Figure 2).

Recipe Retrieval Pipeline Given over 50% of the CSI dish entries in our test set lack detailed publicly available descriptions from sources like Wikipedia, we use the Xiachufang recipe database (Liu et al., 2022) – which contains approximately 1.4 million Chinese monolingual recipes – to retrieve recipes and enhance translation accuracy.

We take inspiration from Translation Studies research, which emphasizes the importance of cooking methods and ingredients in translating Chinese dish names (Amenador and Wang, 2022).

Our retrieval pipeline involves two key stages:

1. **Query and document Construction:** The query is the full Chinese dish name, with the CSI span from previous annotations. We concatenated each recipe name and instructions into a single recipe document.

2. **Filtering and Ranking Recipes:** To filter and rank the recipes, we employ the BM25 algorithm (Robertson et al., 1995), which assigns a score to each word in the Chinese recipe document based on its term frequency and inverse document frequency. For each word in the recipe document, the score is enhanced by applying a weighting factor when the word matches either the dish name (weight = 5) or the CSI span (weight = 3), with an additional multiplier of 3 applied to words within the dish name to prioritize their importance. If there is no exact match for the dish name, the process shifts focus to matching with the CSI span. Additionally, we apply a length penalty to the final score, adjusting it based on the difference between the recipe's length and the average length of all recipes. We select the top-ranked recipe as the final output.

Prompt Strategy: CSI Recipe In this prompting strategy, we use the most relevant recipe returned from the aforementioned search pipeline to aid CSI translation. We provide the name and cooking instructions of the closest-matching recipe while noting that it might not correspond exactly to the given dish but is beneficial as external knowledge since it contains the CSIs to translate.

Prompt Strategy: CSI Explain-then-Translate Inspired by Chain of Thought (CoT) prompting (Wei et al., 2024) and Self-Explanation (Yao et al., 2024), we formulate another prompting strategy that first asks the LLMs to explain the meaning of the CSIs described in the recipe and then generate the translation for the dish. The motivation is to help the LLMs conduct advanced reasoning on the recipe instructions, such as interpreting dish names with CSIs not explicitly defined in the recipe. For example, a recipe might instruct to "cut it first, then stir fry" or note that "it can be very spicy" without explaining the CSI. The LLM's task is to infer the meaning of the CSIs based on the recipe's instructions. The prompt is shown in Figure 2.

5.2 Translation Studies-inspired Prompting

Unlike the conventional prompt engineering used in related work, our second set of prompting strategies differs in that they incorporate *human translation strategies*, inspired by the rich literature in Translation Studies, directly into the design of the prompt. We provide the prompt template for both of these strategies in Figure 2 and complete prompt in Appendix 3 and 4.

Prompt Strategy: Equivalents Using the recipe in §5.1 as external knowledge, we ask the LLMs to produce three translations, each based on a translation strategy (i.e. cultural, functional and descriptive) and **select the best translation**.

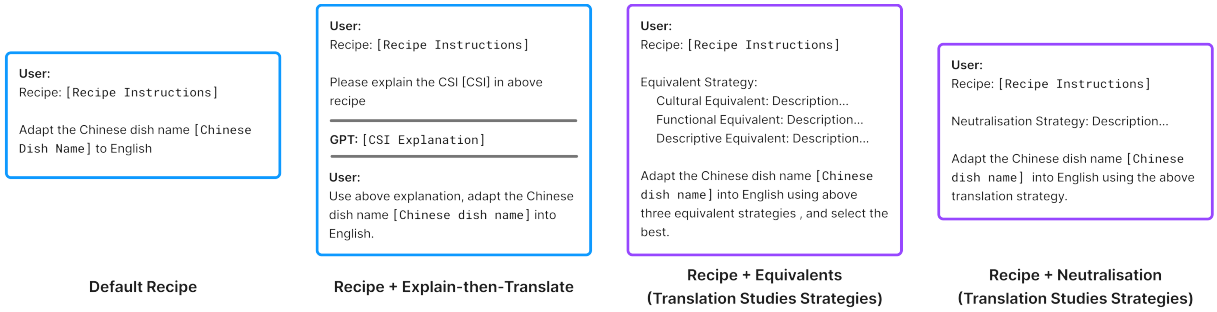

Figure 2: Four adaptation prompt strategies

These translation strategies are inspired by Newmark (1988)'s theories on equivalent translation. We define the equivalent strategies below, providing examples for the reader's understanding:

1. **Cultural Equivalent:** Replacing a CSI in the source text with a term that is culturally relevant and functionally equivalent in the target culture. This strategy aims to evoke the same response in the target audience. (i.e. translating "粽子" as "tamale" in Spanish – given both are traditional wrapped food items made with a starchy substance and fillings, albeit from different cultures.)

2. **Functional Equivalent:** This strategy focuses on the function or purpose of the item (i.e., translating "粽子" as "rice dumpling" to convey the idea of a food made of rice)

3. **Descriptive Equivalent:** Providing a detailed description or explanation of the CSI to convey its meaning and significance. This approach is useful when the CSI is essential for understanding the text but has no equivalent in the target language (i.e. translating "粽子" as "a traditional Chinese sticky rice dumpling wrapped in bamboo leaves")

Prompt Strategy: Neutralisation Another human translation strategy we used in the prompting experiments is neutralisation. Again, we provide recipe information as external knowledge as in the previous strategy and incorporate an explanation of the neutralisation strategy to guide the LLM translation of dish names.

Neutralisation: Using culturally neutral language to describe or explain a cultural word, phrase, or rhetorical expression from the source text. It answers the question, "What is this?" (Amenador and Wang, 2022) by adding information such as ingredients, culinary methods and key characteristics. Compared with the descriptive equivalent strategy, the neutralisation strategy we used for prompt design confines the information used in the translations to ingredients, culinary methods and key characteristics (i.e. translating "粽子" as "sticky rice wrapped in bamboo leaves").

6 Results and Analysis

Our experiments include five parts: 1. Baseline evaluation of MT performance using three models on the ChineseMenuCSI dataset (§6.2); 2. Assessment of CSI span identification accuracy (§6.3); 3. Exploration of main adaptation strategies (§6.4); 4. Exploration of individual equivalents translation strategies for enhancing CSI translation (§6.5); 5. Human evaluation of translation quality on a subset of the dataset (§6.6).

6.1 Experimental Settings

We evaluate the effectiveness of LLM translations for CSIs by comparing various prompting strategies across two SOTA LLMs — GPT-3.5 (gpt-3.5-turbo-0125[4]) and the advanced GPT-4o (gpt-4o-2024-05-13[5]) — against the robust commercial MT system, Google Translate. This approach allows us to assess the strengths of LLM prompting versus a widely used commercial MT.

6.2 Evaluation of CSIs vs Non-CSIs

We use the Stage 1 annotated version of the ChineseMenuCSI dataset (§3.3). This version has a high inter-annotator agreement of 0.91 for the CSI vs non-CSI classification task, with conflicts further resolved using a third annotator. We compare the MT performance using the COMET version wmt22-comet-da (Rei et al., 2022).

While GPT-4o yields major improvements in the CSI translation performance, GPT-3.5 and GPT-4o

[4]GPT-3.5:https://platform.openai.com/docs/models/gpt-3-5-turbo

[5]GPT-4o:https://platform.openai.com/docs/models/gpt-4o

Method	Non-CSIs	CSIs
Google Translate	**74.08**	64.48
GPT-3.5	72.88	64.34
GPT-4o	73.67	**65.97**

Table 3: Comparison of COMET scores across different translation systems for CSI and Non-CSI menu items. The prompt used: "Translate the [Chinese dish name] into English".

	Method	Precision	Recall	F1 Score
CSI-1: Concrete CSIs				
Ours	Combined	64.9	34.0	44.7
	RTT	58.8	28.4	38.3
	CU	38.3	65.3	48.2
	HS	**86.3**	31.2	45.8
GPT	3.5	32.4	**80.4**	46.2
	4o	37.9	67.1	**48.5**
CSI-2: Creatives CSIs				
Ours	Combined	66.1	53.1	58.9
	RTT	63.4	60.1	**61.7**
	CU	35.4	70.4	47.1
	HS	**68.6**	16.4	26.5
GPT	3.5	34.1	**82.0**	48.2
	4o	40.6	73.9	52.4
CSI-3: Abstract CSIs				
Ours	Combined	81.4	68.6	74.4
	RTT	**81.7**	73.6	**77.4**
	CU	43.9	88.4	58.6
	HS	80.0	9.9	17.6
GPT	3.5	50.4	**94.3**	65.7
	4o	59.1	78.9	67.6

Table 4: Evaluation of CSI span identification accuracy by CSI category: precision, recall, and F1 scores.

show worse scores for non-CSIs than Google Translate. These results suggest that Google Translate is particularly strong at translating straightforward, culturally neutral content, likely due to its extensive and diverse training dataset which prioritizes general language accuracy over cultural nuances.

Interestingly, despite Google Translate's general strength in multilingual tasks (Zhu et al., 2024), GPT-4o shows better handling of CSIs, highlighting the benefits of pretraining at scale on diverse corpora from many cultures, as opposed to NMT systems that are typically trained on narrow-domain sentence-level parallel corpora.

6.3 Evaluation of CSI Span Identification

We further assess the capability of different methods to *pinpoint specific CSI spans* – a task we call **CSI Span Identification** – within dish names. This fine-grained analysis is crucial for understanding the elements that contribute to cultural specificity and translation complexity.

Table 4 shows GPT 4o as the best performer in CSI-1 and RTT as the best in other categories. This is likely due to RTT's strength in identifying CSIs, which are often figurative and lack general interpretations. However, the combined metrics only improved RTT's performance in CSI-1, likely because of low recall in HS. Moreover, CU underperforms in CSI-2, as those CSIs typically involve figurative messages in creative combinations of frequent words, which cannot be captured by frequency.

The combined method does not outperform the individual highest method as it requires majority agreement, where a single correct check is insufficient. HS shows high precision but low recall, likely because CSIs often have historical backgrounds, though the inverse is not always true.

6.4 Evaluation of Main Adaptation Strategies

This section examines whether external knowledge, such as recipes, can enhance MT of CSI-rich dish names. Four prompting strategies were tested: a) Default Recipe prompting, b) Recipe + Explain-then-Translate (EtT) prompting, c) Equivalents and d) Neutralisation. For the latter two, which are translation strategies, we try baselines with and without incorporation of recipes, to ablate the dependency on external knowledge. In Table 5, we see that while all our proposed methods yield overall improvements in performance, translation strategy-based methods – that do not involve any external knowledge – yield the largest gains, of up to +4.8 COMET points! Moreover, when recipes are added to these strategies, the gap widens even further, with the maximum gain reaching *as high as +7.87 COMET points* for our best performing *Recipe + Equivalents* strategy. The second trend we note is that the largest gains come from the more complex CSI-2 and CSI-3 categories, indicating the efficacy of our translation theory-inspired methods for translating highly culturally nuanced text. Finally, while the more advanced model GPT-4o naturally yields the largest improvements, we note that we get pretty good results with the far cheaper GPT-3.5 model too, indicating that our methods could be used quite economically.

Revisiting the relatively lower improvements in CSI-1 by examining GPT-generated translations, we find that the LLMs sometimes focus on irrelevant details in the provided recipes. In CSI-1, which involves shorter CSI terms, finding an exact match for dish names is harder, forcing the inclusion of noise in the recipe. For instance, the term "咕噜" *(Sweet and sour)* applies to various dishes

	GPT-3.5				GPT-4o			
	CSI-1	**CSI-2**	**CSI-3**	**Overall**	**CSI-1**	**CSI-2**	**CSI-3**	**Overall**
Baseline	62.68	55.38	43.92	53.33	62.68	55.38	43.92	53.33
Recipe-based Translation								
Recipe	+0.16	-0.90	+3.44	+0.50	-0.08	-3.02	+3.49	+1.93
Recipe + EtT	+1.13	-1.33	+4.92	+1.04	+1.10	+1.61	+4.87	+2.16
Translation Studies Prompting								
Neutralisation	+0.74	+1.15	+3.62	+1.56	+0.46	**+4.84**	+4.29	+3.02
Equivalents	**+1.44**	+3.24	+2.52	+2.38	**+2.34**	+3.89	+0.94	+2.62
Recipe + Translation Studies Prompting								
Recipe + Neutralisation	-1.29	+1.15	**+7.72**	+1.71	+0.09	+3.47	+4.25	+2.34
Recipe + Equivalents	+0.95	**+3.85**	+3.01	**+2.54**	+1.80	+3.24	**+7.87**	**+3.74**

Table 5: COMET score comparisons for GPT-3.5 and GPT-4o using various translation strategies across CSI categories. The overall score is calculated as the average of CSI-1, CSI-2, and CSI-3 scores for each method.

	GPT-3.5				GPT-4o			
	CSI-1	**CSI-2**	**CSI-3**	**Overall**	**CSI-1**	**CSI-2**	**CSI-3**	**Overall**
Baseline	62.68	55.38	43.92	53.33	63.43	54.50	47.50	55.14
Equivalents Strategy Prompting								
Cultural	-0.06	-0.86	+0.99	+0.02	**+0.91**	+0.90	-2.77	-0.32
Descriptive	-6.73	-1.90	+0.93	-2.57	-3.83	+2.62	+2.10	+0.96
Functional	+0.54	+2.69	+0.78	+1.34	+0.06	+3.47	+1.09	+1.54
Recipe + Equivalents Strategy Prompting								
Recipe + Cultural	-2.56	-0.89	-0.06	-1.83	+0.84	+1.73	+0.72	+1.10
Recipe + Descriptive	-8.69	-1.81	+2.29	-2.74	-4.74	**+5.27**	+3.86	+1.46
Recipe + Functional	**+2.27**	**+4.02**	**+2.57**	**+2.95**	-0.96	+2.80	**+7.97**	**+3.27**

Table 6: Ablation study comparing COMET scores for GPT-3.5 and GPT-4o using different equivalent strategies across CSI categories. The overall score is calculated as the average of CSI-1, CSI-2, and CSI-3 scores for each method.

like pork, chicken, or fish, making it difficult to provide the correct ingredient. In contrast, CSI-2 and CSI-3 usually involve longer, more specific phrases like "蚂蚁上树" *(Fried vermicelli with pork)*, making it easier to find an exact recipe match, reduce noise, and majorly improve accuracy.

6.5 Evaluation of Individual Equivalent Strategies

We further perform an ablation analysis of the recipe and individual equivalent strategies, including cultural, descriptive and functional, against the baseline results.

Table 6 shows that for GPT-3.5, the functional equivalent strategy outperforms others, especially when combined with the recipe. For GPT-4o, both descriptive and functional strategies yield better results in CSI-2 and CSI-3, with descriptive strategy excelling in CSI-2 when a recipe is included. In CSI-3, "Recipe + Functional" strategy leads to a significant performance boost of +7.97.

Upon reviewing the translations, both descriptive and functional strategies align well with the gold standards for CSI-2 and CSI-3. However, due to its complexity, the descriptive strategy produces longer translations with trivial details for CSI-3, which is likely to negatively affect COMET scores.

6.6 Human Evaluation

We collect ratings from 10 native Chinese speakers fluent in English, based on the concept of cross-cultural adaptation on a scale of 0 to 10, alongside automatic quantitative metrics. We select the top-performing methods with recipes, as evaluated by COMET in Tables 5 and 6. We then randomly sample 15 entries with perfect agreement from each CSI category (1: Concrete, 2: Creative, 3: Abstract), totalling 45 entries.

The human evaluation results reveal a trend of performance improvement from CSI-1 to CSI-3 in GPT-3.5 and 4o. We use green to highlight cells with major improvements, i.e. over 1 point. Interestingly, for more complex CSIs (i.e. CSI-2 and CSI-3) we have larger improvements. We also observe that these trends align well with COMET trends in Table 5, noting that by both metrics, translation theory prompts yield significantly better results than basic prompting across categories.

Interestingly, human evaluators prefer "*Recipe + Neutralisation*" instead of "Recipe + Equivalent",

	GPT-3.5				GPT-4o			
	CSI-1	**CSI-2**	**CSI-3**	**Overall**	**CSI-1**	**CSI-2**	**CSI-3**	**Overall**
Baseline	6.33	3.88	3.18	4.47	6.22	4.23	3.65	4.70
Recipe-based Translation								
Recipe	-0.93	+0.67	+1.74	+0.49	-0.04	+0.80	+2.28	+1.01
Recipe + Explain-then-Translate	-0.03	+0.60	+1.35	+0.64	+0.43	+0.98	+1.68	+1.03
Recipe + Translation Studies Prompting								
Recipe + Functional	-1.15	+1.10	+1.71	+0.60	+0.75	+1.77	+2.14	+1.05
Recipe + Neutralisation	**+0.37**	**+1.21**	+1.03	+0.62	**+1.32**	**+2.83**	**+3.20**	**+1.95**
Recipe + Equivalent	-0.85	+0.62	**+2.05**	**+0.81**	+0.71	+0.99	+2.38	+1.36

Table 7: Difference in human evaluation of translation quality compared to baseline for different models and strategies across CSI categories. The overall score is calculated as the average of CSI-1, CSI-2, and CSI-3 scores for each method.

the highest in COMET. This preference may stem from the neutralisation definition used in this study, based on the findings of Amenador and Wang (2022). They note that neutralisation is the most common strategy employed by human translators for Chinese names, suggesting a familiarity that could influence the evaluators' preferences towards human-like translation outputs.

Table 8 illustrates the effectiveness of various translation strategies applied to the Chinese dish "三不沾", known for its non-stick quality when served, featuring osmanthus eggs. The full translation examples are provided in Appendix A.3.

Strategy	**Translation**
`Baseline`	Not sticky in three ways
`Equivalent`	Sweet Egg Pastry
`Neutralisation`	Osmanthus Egg Custard

Table 8: Selected Translations from GPT-4o Using Different Translation Strategies with Recipe.

"Sweet Egg Pastry" generated using the Equivalent strategy by GPT-4o, effectively conveys the essence of the dish by focusing on its key ingredients and flavour profile. "Osmanthus Egg Custard," produced through the Neutralisation strategy, is also an accurate translation as it highlights "Osmanthus egg," the main ingredient, and "Custard," indicating the dish's texture. In contrast, the baseline translation "Not sticky in three ways" fails to provide meaningful information about the dish, making it the weakest.

7 Discussion

The CSI categorisation can be applied to wider cultural domains that contain figurative elements. Future research can use this taxonomy to analyze how different translation methods perform on figurativeness and cultural specificity, suggesting a new framework for evaluating CSI translation. This is similar to the evaluating framework in cultural inference, categorising entailment in different levels to better assess an LLM's ability to understand cultural inference (Huang and Yang, 2023).

CSI automatic identification offers a cost-effective approach that outperforms GPT-based prompting in CSI-2 and 3. This method is versatile and applicable to both general and domain-specific CSIs, as it focuses on preserving meaning in translation. It could enhance the quality of translations in a wide variety of domains where maintaining cultural integrity is essential – like literature, media, marketing and cross-cultural communication.

The findings of this paper also demonstrate the effectiveness of prompt strategies inspired by translation studies in overcoming the challenges of translating CSIs, particularly when direct equivalents are lacking across cultures. This approach shows promise for using LLMs with tailored prompts, integrating human translation insights, and translating diverse cultural elements more effectively.

8 Conclusion

In this paper, we introduce the ChineseMenuCSI dataset for CSI-rich dishes and propose a detailed classification in the test set. The results show that LLMs outperformed NMT systems, while NMT is better for Non-CSI translations. Additionally, automatic methods are better than GPT-based prompting at identifying CSIs in most categories.

Incorporating translation studies and recipe details improves LLMs' translation of Chinese dish names. Equivalence strategies, aligned with popular restaurant translations, yield consistently high-quality results, while neutralisation, based on previous analyses, is well-received by evaluators.

Limitations

We acknowledge a few limitations of our study. Firstly, we use COMET as the primary automatic evaluation metric for CSIs. While COMET provides a robust evaluation, assessing cultural awareness may require an even deeper understanding of cultural backgrounds in both source and target languages, which COMET may not fully cover. Currently, in the absence of a metric that can evaluate text-to-text cultural similarity, we use COMET due to its high correlations with human judgment.

Secondly, while we only test zero-shot prompting for translation studies and recipe information, other research, such as Nayak et al. (2024), has demonstrated promising results using few-shot in-context learning strategies, which should also be explored.

Lastly, we only sample 45 menu entries from the test set, which can be relatively small compared to the studies with a larger test set. To achieve more robust and reliable results, increasing the number of human evaluators and the sample size of evaluation entries would be beneficial.

Ethical Considerations

In conducting this research, we adhere to ethical guidelines to ensure the integrity and responsibility of our work. The ChineseMenuCSI dataset is created by scraping publicly available restaurant websites, ensuring that no private or sensitive information is collected. We obtain data in compliance with the terms of use of the websites and anonymise any identifying details of the restaurants. The human annotators involved in this study are fully informed about the nature of the research and provide their consent. We make the dataset available for research purposes under a license that respects the rights of the original content creators.

References

Javier Franco Aixelá. 1996. Culture-specific items in translation. *Translation, power, subversion*, 8:52–78.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Román Álvarez and M Carmen Africa Vidal. 1996. *Translation, power, subversion*, volume 8. Multilingual Matters.

Kate Benedicta Amenador and Zhiwei Wang. 2022. The translation of culture-specific items (csis) in chinese-english food menu corpus: A study of strategies and factors. *SAGE Open*, 12(2):215824402210966.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Alex J Chan, José Luis Redondo García, Fabrizio Silvestri, Colm O'Donnel, and Konstantina Palla. 2023. Harmonizing global voices: Culturally-aware models for enhanced content moderation. *arXiv preprint arXiv:2312.02401*.

Isabelle Chou, Victoria Lei, Defeng Li, and Yuanjian He. 2016. *Translational Ethics from a Cognitive Perspective: A Corpus-Assisted Study on Multiple English-Chinese Translations*, pages 159–173.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Danilo Gallo, Jutta Willamowski, Yada Wisatekaew, Adrien Bruyat, and Antonietta Maria Grasso. 2021. Restaurant menu understanding: Illustrating the need for culturally augmented translation. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20, New York, NY, USA. Association for Computing Machinery.

Federico Garcea, Margherita Martinelli, Maja Miličević Petrović, and Alberto Barrón-Cedeño. 2023. ! translate: When you cannot cook up a translation, explain. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 392–398.

HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. 2023. Bridging background knowledge gaps in translation with automatic explicitation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9718–9735, Singapore. Association for Computational Linguistics.

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

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2022. Deep: Denoising entity pre-training for neural machine translation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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

Vivek Iyer, Edoardo Barba, Alexandra Birch, Jeff Pan, and Roberto Navigli. 2023a. Code-switching with word senses for pretraining in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12889–12901, Singapore. Association for Computational Linguistics.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023b. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore. Association for Computational Linguistics.

Yova Kementchedjhieva, Di Lu, and Joel Tetreault. 2020. The ApposCorpus: a new multilingual, multi-domain dataset for factual appositive generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1989–2003, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024a. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5622–5640, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024b. Indoculture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces. *arXiv preprint arXiv:2404.01854*.

Kai-Chee Lam, Man-Ling Ng, Lay-Hoon Ang, and Radina Mohamad Deli. 2018. Between concrete and abstract: The malaysian chinese way of naming dishes. *International Communication of Chinese Culture*, 5:247–259.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

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

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.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Xiao Liu, Yansong Feng, Jizhi Tang, Chengang Hu, and Dongyan Zhao. 2022. Counterfactual recipe generation: Exploring compositional generalization in a realistic scenario. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7354–7370, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural

alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

P. Newmark. 1988. *A Textbook of Translation*. English language teaching. Prentice-Hall International.

Eugene Albert Nida and Charles R. Taber. 2003. *The Theory and Practice of Translation*. Leiden: Brill.

Valerie Pellatt and Eric Liu. 2010. *Thinking Chinese Translation : A Course in Translation Method: Chinese to English*. London: Routledge.

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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

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

Melanie Soderstrom, Marisa Casillas, Elika Bergelson, Celia Rosemberg, Florencia Alam, Anne S Warlaumont, and John Bunce. 2021. Developing a cross-cultural annotation system and metacorpus for studying infants' real world language experience. *Collabra: Psychology*, 7(1):23445.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Y Tao, O Viberg, RS Baker, and RF Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *arXiv preprint arXiv*, 2311.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. Benchmarking llm-based machine translation on cultural awareness. *Preprint*, arXiv:2305.14328.

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

A Appendix

A.1 Detailed Prompt Teamplates

Prompt Strategy: Recipe + Equivalents

User:

Similiar Recipe: [Recipe Instructions].

Based on the above recipe information, provide three translations for [Chinese dish name] based on the three translation strategies listed below and select the best one:

Cultural Equivalent: Substituting a source language term with a term from the target language that has similar cultural resonance and functionality.

Functional Equivalent: Rendering the source language's meaning, intent, and style into the target language in a culturally appropriate and understandable way. This strategy prioritizes the effect and function of the text in the target culture over a word-for-word translation, ensuring the translation fulfills the same purpose as the original.

Descriptive Equivalent: Providing an in-depth explanation of a term or concept that lacks a straightforward equivalent in the target language. The explanation could include details such as ingredients, culinary method, key characteristics, etc.

Figure 3: Recipe + Equivalents Detailed Prompt

Prompt Strategy: Recipe + Neutralisation

User:

Similiar Recipe: [Recipe Instructions].

Based on the above recipe information, provide a translation for [Chinese dish name] with the following translation strategy:

Menu Description Strategy: This strategy involves using culturally neutral language to describe or explain a cultural word, phrase, or rhetorical expression from the source text (ST). It answers the question, 'What is this?' and is similar to converting a metaphor to its literal meaning. The translations should include the key culinary method, ingredients, and characteristics.

Figure 4: Recipe + Neutralisation Detailed Prompt

A.2 Menu Parser

We develop a heuristic parser to extract dish information from the bilingual menu images by detecting price tags and segmenting the raw text into aligned content. To achieve this, we utilise Google Cloud Vision OCR[6] to extract text and bounding boxes from the menu images. Price tags serve as unique indicators for each dish's content, as we observe that most menus included prices alongside their respective dishes. These price tags are identified using regular expressions, such as "dd.dd" or "£dd.dd".

Given that the position of price tags relative to dish names can vary across menus, we calculate alignment scores based on the cosine similarity and the gap distance between the potentially aligned Chinese and English text and select the alignment with the highest score from all possible combinations. Each entry undergoes manual review to ensure accuracy and errors are corrected before subsequent steps.

[6]Google Cloud Vision OCR: https://cloud.google.com/vision

A.3 Full Examples of Translation Prompts

Strategy	Translation
`Baseline`	Not sticky in three ways
`Reference`	Sweet Egg Pudding
`Recipe`	Three Non-Stick Delicacy: Traditional Imperial Egg Yolk Treat
`Recipe + EtT`	Imperial Non-Stick Egg Delight
`Equivalent`	Sweet, Sticky and Chewy
`Neutralisation`	Non-Sticky Sweet and Savory Egg Custard
`Recipe + Equivalent`	Sweet Egg Pastry
`Recipe + Neutralisation`	Osmanthus Egg Custard

Table 9: Comparison of GPT-4o Translations Across Different Strategies