Bridging Cultures in the Kitchen: A Framework and Benchmark for Cross-Cultural Recipe Retrieval

Tianyi Hu **Maria Maistro** **Daniel Hershcovich**
Department of Computer Science,
University of Copenhagen
tenneyhu@gmail.com, {mm, dh}@di.ku.dk

Abstract

The cross-cultural adaptation of recipes is an important application of identifying and bridging cultural differences in language. The challenge lies in retaining the essence of the original recipe while also aligning with the writing and dietary habits of the target culture. Information Retrieval (IR) offers a way to address the challenge because it retrieves results from the culinary practices of the target culture while maintaining relevance to the original recipe. We introduce a novel task about cross-cultural recipe retrieval and present a unique Chinese-English cross-cultural recipe retrieval benchmark. Our benchmark is manually annotated under limited resource, utilizing various retrieval models to generate a pool of candidate results for manual annotation. The dataset provides retrieval samples that are culturally adapted but textually diverse, presenting greater challenges. We propose CARROT, a plug-and-play cultural-aware recipe information retrieval framework that incorporates cultural-aware query rewriting and re-ranking methods and evaluate it both on our benchmark and intuitive human judgments. The results show that our framework significantly enhances the preservation of the original recipe and its cultural appropriateness for the target culture. We believe these insights will significantly contribute to future research on cultural adaptation.

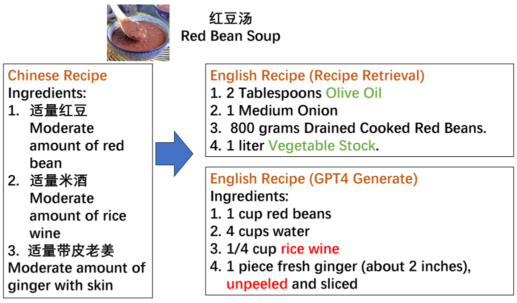
红豆汤
Red Bean Soup
Chinese Recipe
Ingredients:
1. 适量红豆 Moderate amount of red bean
2. 适量米酒 Moderate amount of rice wine
3. 适量带皮老姜 Moderate amount of ginger with skin
English Recipe (Recipe Retrieval)
1. 2 Tablespoons Olive Oil
2. 1 Medium Onion
3. 800 grams Drained Cooked Red Beans.
4. 1 liter Vegetable Stock.
English Recipe (GPT4 Generate)
Ingredients:
1. 1 cup red beans
2. 4 cups water
3. 1/4 cup rice wine
4. 1 piece fresh ginger (about 2 inches), unpeeled and sliced

Figure 1: A cross-cultural recipe adaptation example. The GPT-4 adapted result (Cao et al., 2024), still have some evident shortcomings like using rice wine and unpeeled ginger does not align with culinary practices in English-speaking culture, while the retrieval provided suitable results, includes substitutions of ingredients that align with local culture.

1 Introduction

Cooking recipes are key tools in culinary culture (Borghini, 2015), which largely varies by culture and language (Albala, 2012). For example, geographical conditions significantly affect ingredients availability while culinary history shapes people's taste preferences. Food choice is a complicated behavior (Köster, 2009) and it is highly associated with socio-cultural factors (Rozin, 1996). The familiarity of food products is positively associated to sensory liking (Torrico et al., 2019).

Recognizing and adapting cultural differences presents both a significant importance and a challenge (Hershcovich et al., 2022). Merely translating recipes can lead to both semantic and cultural mismatches (Yamakata et al., 2017; Zhang et al., 2024). As shown in Figure 1, even GPT-4, a powerful Large Language Models (LLMs) and a state-of-the-art (SOTA) model in cross-cultural recipe adaption (Cao et al., 2024), still makes obvious mistakes when adapting recipes from one culture to another, e.g., the selection of ingredients and tools are not commonly used or the flavors do not align with the preferences in the target culture. We propose to use Information Retrieval (IR) methods to address the issue because compared to generative models, retrieved recipes from a target culture corpus naturally align more closely with the target culture in flavor, ingredients and tools.

Nevertheless, cross-cultural recipe IR is a challenging task due to the existing linguistic and cultural gap between the source and target. Besides

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1068–1080
November 12-16, 2024 ©2024 Association for Computational Linguistics

the challenges posed by the intrinsic gap between different languages (Zhang et al., 2022), an even bigger challenge is the textual discrepancies caused by cultural differences in dietary habits, naming conventions, and food-related knowledge, which complicate the task. We identify three non-trivial challenges related to cross-cultural recipe retrieval:

Relevance Assessment for Cross-Cultural Recipes Retrieval Due to cultural variations in ingredients, seasonings, and cooking methods, assessing the relevance of cross-cultural recipe pairs is complex and challenging, needing clear guidelines to standardize relevance assessments.

Culture-Aware Framework to Bridge Cultural Gaps Current IR models lack awareness of the significant cultural gaps that exist across diverse culinary traditions, retrieving recipes that are textually similar but actually quite different.[1]

Benchmark of Cross-Cultural Recipe Retrieval No currently publicly available dataset[2] can be used as a benchmark to evaluate the performance of different retrieval models and to understand how cultural differences present challenges to our task.

Our contributions to tackle these challenges are:[3]

1. We introduce the novel cross-cultural recipes retrieval task. We provide assessment guidelines for cross-culture recipes relevance judgement, with specific criteria and examples.

2. We propose CARROT, a plug-and-play cultural-aware recipe IR framework, and demonstrate that it offers better relevance compared to the results of previous retrieval models and better consistency and cultural appropriateness to the results generated by LLMs on Chinese-English recipe cultural adaption.

3. Focusing on recipes in Chinese and English, we design and annotate a cross-cultural recipe retrieval dataset. It has many challenging samples like cultural differences leading to significant textual discrepancies in matched recipes.

2 Related Work

Cultural and Recipe Adaptation Cultural adaptation aims at changing the text's style by the attributes of culture while maintaining its original meaning, it involves common ground, values and aboutness (Hershcovich et al., 2022). Recipe adaptation is an important application of cultural adaptation, Liu et al. (2022) demonstrated that recipe adaptation is a challenging task. Although language models can generate fluent recipes, they struggle to use culinary knowledge in a compositional way, such as adjusting cooking actions related to the changing ingredients. Palta and Rudinger (2023) and Zhou et al. (2024) underscore the complexity of integrating cultural understanding into LLMs, particularly in the culinary domain. Cao et al. (2024) propose the cross-cultural recipes adaptation task and show that prompting LLMs for recipe generation is the SOTA method for this task. They build a recipe adaptation dataset automatically using an IR model to match recipes. However, their purpose is not to propose a novel IR model—an off-the-shelf standard IR model is used and is not evaluated with respect to the retrieval task.

Recipe Retrieval Works in recipe retrieval primarily focus on cross-modal recipe retrieval (Lien et al., 2020; Salvador et al., 2021), retrieving recipes by both text and images. Takiguchi et al. (2021) introduce a recipe retrieval model for Japan's largest recipe sharing service. Their model is trained and evaluated with online search logs. These works are not primarily aimed at cross-cultural scenarios, and they use online behavior logs as datasets, whereas our work requires the use of manually annotated samples.

LLMs for Information Retrieval The emergence of LLMs has profoundly impacted IR due to their remarkable abilities in language understanding. LLMs for query rewriting have been widely applied to various retrieval issues which have vocabulary mismatches between queries and documents (Zhu et al., 2023). For example, Tang et al. (2023) propose a prompt-based input reformulation method to tackle the problem of inputs in legal case retrieval that often contain redundant and noisy information. LLMs are also widely used for reranking. Even without fine-tuning, they have been proven to possess strong ranking capabilities (Zhu et al., 2023), even superior to state-of-the-art supervised methods on popular IR benchmarks (Sun et al., 2023). We adapt the existing work for cross-cultural recipe retrieval to address the unique

[1] See examples in Figure 3.

[2] Previous work used IR methods only to construct datasets, but these cannot serve as evaluation datasets for IR.

[3] The code and dataset are available at https://github.com/TenneyHu/CARROT

challenges within the domain.

3 Cross-Cultural Recipe Retrieval Task

We define the task of cross-cultural recipe retrieval with the source recipe as query and recipes from the target culture as documents. For a pair consisting of different cultural recipes (q, d), which represents a pair of one query and one document, we assess relevance with a three-point scale: 0 (Not Match), 1 (Partial Match), and 2 (Exact Match), the three-point scale levels are a common choice for relevance assessment (Kekäläinen and Järvelin, 2002).

For Exact Match recipes, the differences should not exceed the necessary range of cultural adaptation, such as making local adjustments with similar ingredients and flavors according to the target culture. Partial Match recipes have similarities in some aspects of ingredients and flavors, offering reference value. They should be in the same dish category (e.g., main courses, desserts, beverages). If the above conditions are not met, the two recipes will be deemed Not Match. We provide specific criteria and examples of relevance assessment in the Appendix A.

We briefly summarize **three main challenges** in the cross-cultural recipe retrieval task:

C1: Is Recipe Title the Best Retrieval Query? Semantic Gaps Caused by Cultural Differences The recipe title is often used as the query (Cao et al., 2024) to retrieve recipes from the target culture, as the title usually encapsulates the essence of the recipe. However, due to language and cultural differences, it forms a semantic gap between the source and target recipe titles in different cultures. These differences include:

Naming Conventions Recipes are typically named after the main ingredients and cooking methods in English-speaking cultures, whereas Chinese cuisine may name dishes after the inventor or origin city, such as *Kung Pao Chicken*.

Culinary Cultures Cultural differences require substituting original ingredients and cooking methods with more locally common alternatives. These changes are also reflected in textual variations between recipe titles. For instance, *Stir-fried Taro*[4] could be adapted to *Stir-fried Potatoes*.

Food-related Common Sense Recipes implicitly contain food-related knowledge that might be common in one culture but unknown in another, e.g., in Chinese cuisine, 地三鲜 (literally, *Three Fresh Ingredients in the Earth*) refers to a dish made with potatoes, eggplants, and green peppers. The specific ingredients represented here are cultural common sense in China but may be challenging for users in other cultures.

C2: Lack of Matching Recipe Samples Considering the high cost of collecting a large-scale manually annotated dataset and the lack of a publicly available dataset, training models is challenging.

C3: Beyond Relevance: Cultural Adaptation in Ranking Current retrieval models primarily rank based on relevance; however, in cross-cultural recipe retrieval, cultural appropriateness is also an important factor to consider in ranking.

4 CARROT: A Cultural-Aware Recipe Retrieval Framework

We propose a framework **CARROT**: **C**ultural-**A**ware **R**ecipe **R**etrieval **T**ool, as shown in Figure 2, a plug-and-play model combining prompt-based LLMs and IR methods, to address the additional challenges posed by cultural differences.[5] Specifically, to address C1 in Section 3, we introduced query rewriting by LLMs. To address C2, we introduce a plug-and-play framework (no additional fine-tuning required). To address C3, we design an additional re-ranking stage.

Query Processing Processing can be divided into translating the query and rewriting the query. The task differs from a general recipe search because the query is not a user-written set of keywords, but a source recipe and title serves as a good summary of the relevant content for the search. So for a Chinese recipe, we first automatically translate its title into English as the original query. We also utilize LLMs for two rewriting tasks. Both the rewritten and original queries are used for retrieval to further enhance the system's robustness, as each query may experience some semantic errors.

Recipe Title Generation Task Inspired by doc2query (Nogueira et al., 2019), we mask the original recipe titles and then prompt LLMs with the ingredients and cooking steps in the recipe to

[4] Taro is a staple root vegetable in Chinese cuisine, not readily available in Western countries.

[5] The prompt used here is shown in Appendix B.

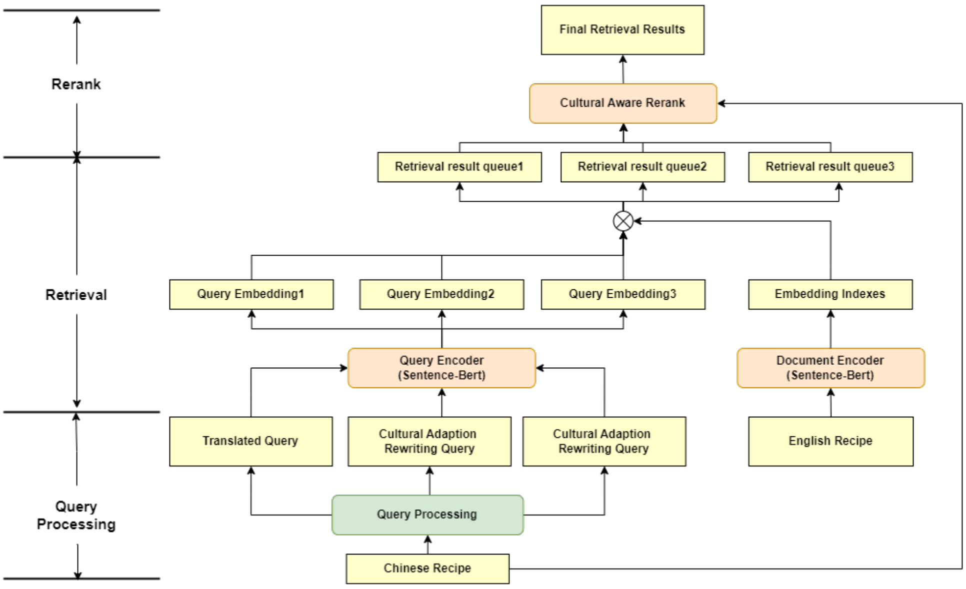

Figure 2: Framework of CARROT, including three stages: Using LLMs for query rewriting, retrieval and Re-ranking based on cultural adaptability and relevance. Different queries will use different embeddings for retrieval and obtain different retrieval result lists. They will be merged during the re-rank stage.

regenerate a title. We believe such generated titles can eliminate interference caused by inappropriate original titles, e.g., users may submit attention-grabbing but non-standard recipe titles, or titles that use personal names or historical references.

Recipe Title Cultural Adaption Task We also prompt LLMs to directly rewrite an English recipe title based on the Chinese recipe title, making it more in line with the writing conventions of recipes in the target culture.

Retrieval Considering millions of recipes in the target culture, we choose a bi-encoder structure to efficiently retrieve the recipes of the target culture. We perform retrieval for each query individually, retaining the top 10 results of each query.

Re-ranking A complex re-ranking model can better understand the implicit culinary cultural knowledge and be more effective, considering factors of cultural matching in ranking. We prompt LLMs to rank the results based on relevance and prioritize recipes that are more aligned with the target culture when the relevance level is the same. Considering the potential issues of using LLMs as unsupervised rerankers, such as limitations in context length and more positional bias compared to traditional models (Zhu et al., 2023), we avoided ranking the retrieval results at once. Instead, we performed multiple rounds of ranking or combined LLMs with other rerankers (Xiao et al., 2023).

5 Cross-Cultural Recipe Retrieval Dataset

5.1 Recipe Corpora

We source recipes from two monolingual corpora: RecipeNLG (Bień et al., 2020) and XiaChuFang (Liu et al., 2022). RecipeNLG has over two million English cooking recipes and XiaChuFang consists of more than 1.5 million Chinese recipes from a Chinese recipe website.[6] We use the title, ingredients, and cooking steps from each corpus. These two corpora are independent and monolingual. Therefore, we use the Chinese recipe corpus as the source and annotate the relevance of recipes from the English corpus.

5.2 Dataset Construction

Our work draws inspiration from the *Cultural-Recipes* Dataset (Cao et al., 2024), which, however, lacks an evaluation of the retrieval methods and relies on a single method. This introduces potential biases to the dataset, omitting difficult-to-recall positive examples and challenging negative examples, which are vital for robust IR (Zhan et al., 2021). Another challenge is the limitation of annotated resources. The corpora in Section 5.1 contain millions of recipes, the majority of which are irrelevant for a given query.

To address these gaps, we devise manually an-

[6] xiachufang.com

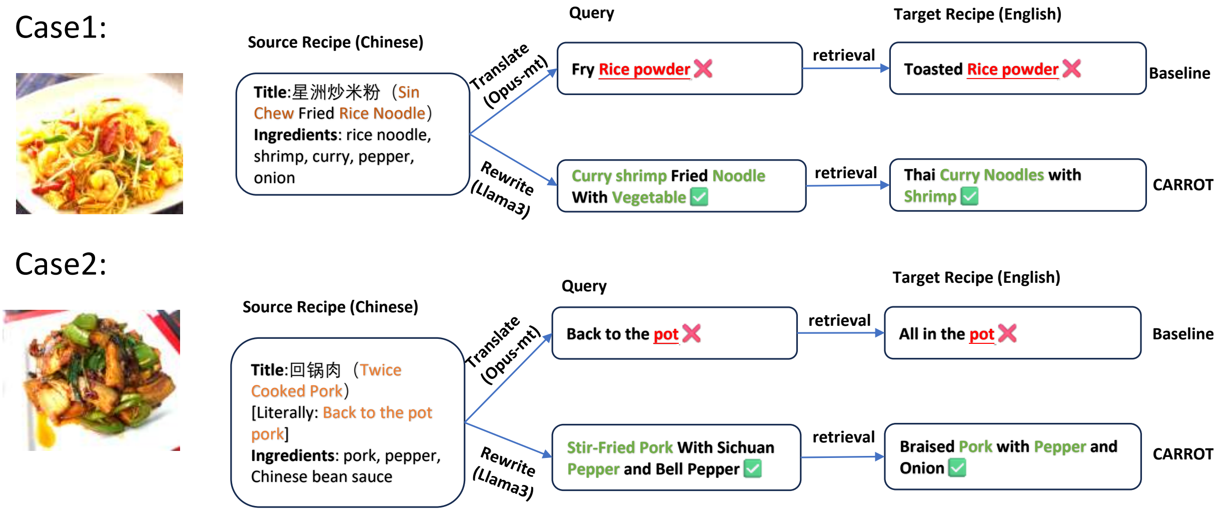
Case1:
Source Recipe (Chinese)
Title:星洲炒米粉（Sin Chew Fried Rice Noodle）
Ingredients: rice noodle, shrimp, curry, pepper, onion
Translate (Opus-mt)
Rewrite (Llama3)
Query
Fry Rice powder ✗
Curry shrimp Fried Noodle With Vegetable ✓
retrieval
retrieval
Target Recipe (English)
Toasted Rice powder ✗
Thai Curry Noodles with Shrimp ✓
Baseline
CARROT
Case2:
Source Recipe (Chinese)
Title:回锅肉（Twice Cooked Pork）
[Literally: Back to the pot pork]
Ingredients: pork, pepper, Chinese bean sauce
Translate (Opus-mt)
Rewrite (Llama3)
Query
Back to the pot ✗
Stir-Fried Pork With Sichuan Pepper and Bell Pepper ✓
retrieval
retrieval
Target Recipe (English)
All in the pot ✗
Braised Pork with Pepper and Onion ✓
Baseline
CARROT

Figure 3: Case Study with two examples, comparing our framework (CARROT) with the baseline (machine translation and MPNet). In the first example, sin chew, refers to Singapore, denotes a curry flavor style and rice noodles are not commonly found in Western countries, the translated query changes it to rice powder, a semantically similar but distinctly different food, while our framework solves these two issues using curry and noodles to adapt the recipe. In the second example, twice-cooked pork is a unique Chinese dish containing specific knowledge. The translated query back to the pot is literally similar but does not describe the flavor and ingredients. Our framework uses the ingredients pork & pepper and cooking methods to explain the dish, making it more conducive to retrieval.

notated samples instead of automatically matched samples and create a candidate pool by multiple retrieval methods for annotation. We randomly pick source recipes from Chinese recipe corpora and build a candidate pool by target culture recipes corpora using multiple retrieval methods. We randomly select recipe samples for manual annotation within the candidate pool. We present statistical information about the dataset in Table 1. For about 83.7% of the queries, the dataset provides at least one document that is an exact match.

The dataset is independently annotated by two voluntary annotators whose native language is Chinese and who are fluent in English. They are also familiar with the culinary practices of both Chinese and English-speaking cultures. The annotators follow the instructions in Appendix A.

Build Candidate Pool We employ a depth-10 pooling strategy to annotate the dataset, which is a standard procedure in IR (Pavlu and Aslam, 2007). Compared to random sampling, using a pooling strategy provides more relevant rather than randomly irrelevant samples. Additionally, compared to annotating the dataset using results from a single retrieval method, the dataset's sources are more diverse and less biased, enhancing the reusability of the dataset. The depth is set to 10 based on the trade-off between the reusability of the dataset and the available annotation resources. We employ four types of retrieval methods to construct the candidate pool:

Attribute	Information
Recipe Corpora: # Recipes	
English corpus size	2 million+
Chinese corpus size	1.5 million+
Dataset Size	
# Queries	98
# Query & Document Pairs	1517
# Average Pairs Per Query	15.5
Annotators	
# Annotators	2
Cohen's Kappa Agreement	0.67
Candidate Pool	
Pool Depth	10
Total Pool size	70–90
Dataset Distribution	
Exact Match Pairs	33.3%
Partial Match Pairs	56.2%
Not Match Pairs	10.5%

Table 1: Statistical Information of Recipe Corpora, Dataset size, Annotator, Candidate Pool and Dataset Distribution in the IR Dataset.

Basic Method We use the Chinese title translated to English as query for two independent SOTA vector-based retrieval models, MPNet sentence-transformer (Song et al., 2020; Reimers and Gurevych, 2019) and ColBERT (Khattab and Zaharia, 2020).

Content-Based Retrieval Compared to only using the titles in the basic method, considering incompleteness of information in titles, we also use the content-based retrieval by title appended with ingredients.[7]

Multilingual Retrieval We also use multilingual sentence-BERT model (Reimers and Gurevych, 2020a) to retrieve instead of translating the query. We directly use untranslated Chinese recipe titles to retrieve English recipes.

Query Rewriting We use both of two rewriting methods in Section 4 and also manually rewrite an alternative title on 48% of the recipes, which are considered to have better alternative queries by manual checking.

6 Experiments

We describe our recipe retrieval experiments and results, using the dataset introduced in Section 5 and *CulturalRecipes* (Cao et al., 2024), a manually annotated cross-cultural recipe adaptation dataset, to compare the results with LLMs generated.

6.1 Metrics

IR Evaluation We use common metrics in IR, including **nDCG@10**, Precision@10 (**P@10**), Precision@1 (**P@1**), Recall@10 (**R@10**), and mAP@10(**mAP@10**).[8] Different IR metrics can contribute to the results in various ways, Precision ensures that the most relevant recipes appear at the top, while NDCG evaluates the overall quality and order of the list. Recall is crucial for capturing all relevant options, providing flexibility for further refinement of recipe rankings based on users' specific dietary preferences. These comprehensive metrics offer references for various downstream applications of recipe retrieval.

Due to limited annotation resources and the pooling strategy, our annotations are incomplete. Following previous work (Sakai and Kando, 2008), in Section 6.3, we only present results for evaluation with condensed lists (non-labelled samples are discarded). Additionally, we include evaluation with full lists (non-labelled samples are considered non-related) results in the Appendix D. The conclusions of the two experiments are similar.

Recipe Adaptation Evaluation We evaluate the IR results using metrics from the recipe cultural adaptation task (Cao et al., 2024) to obtain end-to-end adaptation performance and directly compare the results with those generated by LLMs. We first use reference-based automatic metrics. Since these are not always reliable for subjective tasks, we also perform manual evaluation method with a 7-scale rating in four different aspects.

Reference Based Automatic Evaluation To evaluate the similarity between the retrieved and reference recipes, we use three overlap-based metrics: BLEU (Papineni et al., 2002), ChrF (Popović, 2015), ROUGE-L (Lin, 2004) and one representation-based metric: BERTScore (Zhang et al., 2019).

Human Evaluation The same annotators as in Section 5.2 perform manual evaluation on four criteria in cross-cultural recipes adaptation, adopted from Cao et al. (2024):

Grammar (GRA): The results are grammatically correct and fluent.

Consistency (CON): The results include a complete and detailed title, ingredients and steps, facilitating users to cook according to the recipe.

Preservation (PRE): The results retain the original ingredients and flavors of the source recipe.

Cultural Appropriateness (CUL): The results conform to the dietary habits and recipe writing conventions of the target culture.

Each dimension is rated on a 7-point scale and a higher score indicates superior performance. In addition, we also annotate the 3-scale relevance of recipe retrieval results and computed the Exact match precision at the first position (P@1).

We use Krippendorff's alpha (Vogel et al., 2020) to measure the annotation agreements, which results in 0.79, 0.65, 0.61, 0.82, 0.42 for Relevance Score, Grammar, Consistency, Preservation, and Cultural Appropriateness respectively, indicating substantial agreement between the annotators on most aspects, but a high degree of subjectivity in the understanding of Cultural Appropriateness.

[7] We do not use cooking steps because they are too lengthy and contain little information useful for retrieval.

[8] In Precision, Recall and mAP, only exact matches are considered relevant results while partial matches are treated as irelevant results.

Method	nDCG@10	R@10	mAP@10	P@10	P@1
Basic Retrieval Model					
ColBERT	0.237	12.99	11.99	7.96	5.10
ColBERT Content-based	0.191	6.95	7.41	3.98	5.10
Sentence-transformer Content-based	0.194	9.25	11.77	6.02	6.12
Sentence-transformer	0.298	20.73	20.57	11.63	10.20
Multilingual Sentence-transformer	0.227	19.30	17.96	8.27	13.27
Query Rewrite					
Llama3 Recipe Title Cultural Adaption	0.303	35.67	**27.50**	13.27	**15.31**
Llama3 Recipe Title Generated	0.258	21.25	15.46	7.96	10.20
Reranking					
Sentence-transformer + Llama3 Re-rank	0.305	20.98	21.14	11.63	**15.31**
CARROT (Rewriting + Re-ranking)					
CARROT-Llama3	**0.346**	**37.05**	25.97	**15.71**	**15.31**

Table 2: Evaluation on the cross-cultural recipe retrieval dataset, higher scores indicate better performance on all metrics. Please refer to Section 5.2 for details on the basic retrieval model, and for query rewrite and re-rank in section 4. **Bold** indicates the best performance across all method, underlined indicates best performance across all basic retrieval model. The results show both recipe title cultural adaptation and re-ranking improve relevance.

6.2 Experimental Setup

We represent a recipe as a concatenation of title, ingredients and steps. For constructing the cross-cultural recipe retrieval dataset, we translate Chinese recipe to English by opus-mt models (Tiedemann and Thottingal, 2020), and retrieve English recipes by MPNet sentence-transformer (Song et al., 2020) and ColBERT (Santhanam et al., 2021; Khattab and Zaharia, 2020). We also explore multilingual sentence-transformer (Reimers and Gurevych, 2020b). In the CARROT framework, we set MPNet as the default retrieval model. We explore the performance of using only re-ranking or using only a specific type of query rewriting and various LLMs which are trained on both Chinese and English to enhance the performance of the framework. These models include: Llama3-7B (AI@Meta, 2024), Qwen1.5-7B (Bai et al., 2023) and BAICHUAN2-7B (Baichuan, 2023), the leading Chinese open-source LLMs models[9] and among them Llama3 is currently the best-performing Chinese LLMs under 10B parameters. All the above models are run with default hyper-parameters.

The annotator information is the same with annotators in Section 5. The prompts we use are in Appendix B. We list the versions of the models used in Appendix C.

6.3 Experimental Results

Information Retrieval Results Table 2 shows the results on cross-cultural recipe retrieval dataset in Section 5. Within the basic retrieval models, the Sentence-transformer based on translated titles achieved best overall performance, it is also the reason we use MPNet as the default retrieval model in the CARROT framework. We can find the cultural adaptation rewriting shows better relevance performance compared to translated titles, which proves Chinese recipe titles are not entirely suitable for the naming conventions of English recipes, as well as the effectiveness of the rewriting approach. The CARROT-Llama3 achieve the best performance on nDCG, R@10, P@1, P@10 and the second best performance on mAP@10, demonstrates the strong performance of our framework in this task.

Recipe Adaptation Results Table 3 shows the performance on reference based automatic evaluation and human evaluation. We find that generation methods outperform retrieval methods on the ROUGE-L, BertScore, P@1, Preservation metrics, indicating that the generation method has better relevance and is more faithful to the source recipes, while retrieval methods achieved better results in Consistency and Cultural Appropriateness.[10] The Kendall correlation between P@1 relevance metric and Preservation is 0.73, which indicates that Preservation can also effectively reflect the rele-

[9] According to https://github.com/jeinlee1991/chinese-llm-benchmark.

[10] Further explanations on how our framework enhances them in Section 7.

Methods	BLEU	Chrf	ROUGE-L	BertScore	P@1	GRA	CON	PRE	CUL
Baseline									
Translated Title (opus-mt-zh-en)	20.17	31.78	17.46	59.43	0.64	5.96	5.2	4.2	5.92
Rewrite Only									
Llama3 Recipe Title Generated	**22.14**	**43.38***	18.52	60.70	0.68	**6.0**	5.52*	4.32	**6.2***
Llama3 Recipe Title Cultural Adaption	20.06	38.54*	19.18	60.29	0.8	**6.0**	5.32	4.92*	6.16*
Re-ranking Only									
Translated Title + Llama3 Re-rank	14.25	31.03	17.91	59.85	0.72	5.96	5.48	4.32	6.0
Carrot (Rewriting + Re-ranking)									
CARROT-Llama3	15.90	38.45*	**19.46**	**61.12**	**0.92**	**6.0**	**5.64***	**5.04***	6.16*
CARROT-BAICHUAN	21.86	34.65	17.49	59.45	0.72	**6.0**	5.32	4.4	5.92
CARROT-QWEN	13.44	38.19*	16.31	59.34	0.84	5.96	5.4	4.6	5.92
*Llama3-Generation	19.60	40.26*	32.10*	66.41*	1.0	5.92	5.17	6.04*	5.0

Table 3: Automatic and Human Recipe Adaptation Evaluation on *CulturalRecipes* Dataset: the first four metrics automatically calculated based on reference and the next five metrics are evaluated by human, higher scores indicate better performance on all metrics. We set MPNet as retrieval model here. **Bold** indicates best performance across all retrieval models, and underlined indicates that the generative model outperformed the best retrieval models in this metric. Better results than *Baseline* with significance difference for p < 0.05 by t-test is indicated by *. It shows generation methods outperform in relevance while retrieval is better in consistency and cultural appropriateness.

vance between the results and the source recipes.

Within the retrieval methods, compared to the translated title, both query rewriting methods and re-ranking significantly improved relevance related metrics. The CARROT framework with Llama3 outperforms CARROT with the other two Chinese LLMs, Qwen and Baichuan, highlighting the strong performance of the Llama3 model on cross-lingual tasks. The CARROT-Llama3 achieved the best performance on ROUGE-L, BertScore, P@1, Preservation and Consistency metrics and near-optimal performance on Cultural Appropriateness metrics within the retrieval methods. It demonstrates the strong performance of our framework in the cross-cultural recipe adaptation task.

Case Study We select some cases to intuitively compare the result of using the CARROT framework versus the baseline, just using the translated recipe title and a bi-encoder MPNet model (Song et al., 2020), shown in Figure 3. The results shows machine translation title used as a query can lead to irrelevant search results due to cultural differences, but our CARROT framework addresses this issue by changing the way recipes are named and substituting ingredients.

7 Discussion

The previous SOTA generation method in the task of cross-cultural recipe adaptation shows better relevance. However, retrieval methods are superior in consistency and cultural appropriateness. Our work is the first to highlight the potential issues in using LLM-generated content for recipes, as well as the potential advantages of using IR methods for cultural adaptation. We will illustrate through specific examples how retrieval methods may have advantages over generation methods in these aspects.

Consistency Consistency mainly reflects the quality and reliability of the recipes, which determines whether people can successfully cook according to such recipes. The recipes retrieved are based on real human culinary practices, but recipes generated by LLMs, despite being textually close to user created recipes, still contain hallucinations, leading to not truly instructive texts for human cooking. For example, Llama3 generates the cooking steps of *Braised Beef with Potato* as:

```
4. ... covered for 1 hour or until the beef is
tender.
5. Remove the pot from the heat and discard
```

The discarding in the final step does not align with general culinary understanding and this issue does not exist in the retrieval results.

Cultural Appropriateness The generation method tends to preserve the original flavors, making only necessary changes such as measurement units. In contrast, the retrieval-based method makes more substantial modifications to the ingredients and flavors to better adapt to the culture. For example, for *Salted Baked Chicken* would be adapted to *Salt-Rubbed Roast Chicken with Lemon & Thyme* with the addition of **lemon** and **thyme** to better suit local preferences.

Diversity The retrieval models can find results with significant differences in ingredients and flavors, providing a broader range of references. For example, there are more than 5 different main ingredient combinations in the recipe *red bean soup* top 10 retrieval results by the CARROT framework, with manually highlighted specific ingredients used.

```
1.Dried Red Kidney Beans, Butter, Onion
2.Drained Cooked Red Beans, Olive Oil, Onion
3.Red Beans,Pork,Sprig of Thyme, Canned Tomato
4.Canned Red Kidney Beans, Garlic Bud, Sausage
5.Red Kidney Beans, Celery Stalk, Onion, Carrot
```

8 Conclusion

In this paper, we propose a novel task of cross-cultural recipe retrieval, we have manually annotated a challenging and representative benchmark. Furthermore, we introduce CARROT, a cultural-aware recipe retrieval framework that utilizes LLMs to rewrite and re-rank, thereby bridging the cultural differences in recipes between two distinct cultures. Our approach has robust performance on both our proposed dataset and cultural recipe adaption dataset. We also discuss the advantages of using IR methods for cultural adaptation of recipes versus direct generation using LLMs. We believe our work offers a new perspective on cultural adaptation.

Limitations

Our study presents a benchmark and framework for cross-cultural recipe retrieval, but we acknowledge certain limitations within our study, which may warrant further exploration:

Large scale manual evaluation While our study conducts a small-scale benchmark to evaluate the performance of IR models, the small-scale dataset limits the accuracy of evaluating some IR methods, especially those that significantly differ from the dataset constructed in our work. In an ideal scenario, the benchmark necessitates a large-scale human evaluation of different backgrounds and cultures. Such a large-scale benchmark would prove challenging owing to the significant resources to achieve.

Coverage of recipes from different cultures Although we believe that our proposed framework can be extended to other languages and cultural backgrounds, due to limitations in resources and the background of annotators, we conducted our research using only the Chinese-English example. Ideally, the benchmark and experiments could be extended to include other languages and cultural backgrounds. Studying other culinary cultures might also bring new inspiration to our methods.

Fine-tuning the retrieval model Due to limitations in annotation resources, we directly used the current popular retrieval models without fine-tuning them. Recipe retrieval is a specialized task that requires retrieval models to learn language and knowledge in the food domain. Therefore, ideally, collecting relevance data specific to recipes and fine-tuning the models would enhance the overall performance of the framework.

Acknowledgements

This research was co-funded by the Villum and Velux Foundations Algorithms, Data and Democracy (ADD) grant. Thanks to the anonymous reviewers and action editors for their helpful feedback. The authors express their gratitude to Yong Cao for his assistance with the dataset collection process. We also extend our sincere gratitude to Jingyi Zheng for his valuable feedback on the paper.

References

AI@Meta. 2024. Llama 3 model card.

Ken Albala. 2012. *Three World Cuisines: Italian, Mexican, Chinese*. Rowman Altamira.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. RecipeNLG: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on*

Natural Language Generation, pages 22–28, Dublin, Ireland. Association for Computational Linguistics.

Andrea Borghini. 2015. What is a recipe? *Journal of Agricultural and Environmental Ethics*, 28:719–738.

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.

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.

Jaana Kekäläinen and Kalervo Järvelin. 2002. Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Egon P Köster. 2009. Diversity in the determinants of food choice: A psychological perspective. *Food quality and preference*, 20(2):70–82.

Yen-Chieh Lien, Hamed Zamani, and W Bruce Croft. 2020. Recipe retrieval with visual query of ingredients. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1565–1568.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

V Pavlu and J Aslam. 2007. A practical sampling strategy for efficient retrieval evaluation. *College of Computer and Information Science, Northeastern University*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2020a. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Nils Reimers and Iryna Gurevych. 2020b. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Paul Rozin. 1996. The socio-cultural context of eating and food choice. In *Food choice, acceptance and consumption*, pages 83–104. Springer.

Tetsuya Sakai and Noriko Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11:447–470.

Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. 2021. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15475–15484.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in* *neural information processing systems*, 33:16857–16867.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Kentaro Takiguchi, Mikhail Fain, Niall Twomey, and Luis M Vaquero. 2021. Evaluation of field-aware neural ranking models for recipe search. *arXiv preprint arXiv:2105.05710*.

Yanran Tang, Ruihong Qiu, and Xue Li. 2023. Prompt-based effective input reformulation for legal case retrieval. In *Australasian Database Conference*, pages 87–100. Springer.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Damir Dennis Torrico, Sigfredo Fuentes, Claudia Gonzalez Viejo, Hollis Ashman, and Frank R Dunshea. 2019. Cross-cultural effects of food product familiarity on sensory acceptability and non-invasive physiological responses of consumers. *Food research international*, 115:439–450.

Carl Vogel, Maria Koutsombogera, and Rachel Costello. 2020. Analyzing likert scale inter-annotator disagreement. *Neural approaches to dynamics of signal exchanges*, pages 383–393.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Yoko Yamakata, John Carroll, and Shinsuke Mori. 2017. A comparison of cooking recipe named entities between japanese and english. In *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with The 2017 International Joint Conference on Artificial Intelligence*, pages 7–12.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512.

Fuwei Zhang, Zhao Zhang, Xiang Ao, Dehong Gao, Fuzhen Zhuang, Yi Wei, and Qing He. 2022. Mind the gap: Cross-lingual information retrieval with hierarchical knowledge enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4345–4353.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024. Cultural adaptation of menus: A fine-grained approach. *arXiv preprint arXiv:2408.13534*.

Li Zhou, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. 2024. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. *arXiv preprint arXiv:2404.06833*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Specific Criteria and Examples of Cross-Cultural Recipe Retrieval Task

Criteria of Exact Match A recipe pair that is an exact match should fully satisfy the user's needs for seeking a recipe that is both similar to the source recipe and in line with the target culture. An exact match recipe pair (q, d), should meet one of the following two criteria:

1. The dishes in the two recipes are **consistent**, which means they maintain high similarity in the main ingredients and flavors.

2. The dishes in the two recipes are similar, where differences must reflect cultural differences between source and target

For the first criteria, two dishes are considered **consistent** if they:

1. Use the same main ingredients

2. Employ similar preparation methods

3. Result in a similar taste

For example:

- *Mapo Tofu (Spicy Tofu)* and *chili con carne* are **inconsistent**, even though their flavors are similar, because their main ingredients are different.

- *Spicy fried cabbage* and *Cabbage Soup* are **inconsistent** because they have significant differences in flavor.

Method	**nDCG@10**	**R@10**	**mAP@10**	**P@10**	**P@1**
Basic Retrieval Model					
ColBERT	0.133	12.36	10.58	7.03	4.50
ColBERT Content-based	0.101	6.01	6.54	3.51	4.50
Sentence-transformer Content-based	0.104	8.12	10.39	5.32	5.41
Sentence-transformer	0.182	20.47	18.16	10.27	9.01
Multilingual Sentence-transformer	0.132	16.71	15.86	7.30	11.71
Query Rewrite					
Llama3 Recipe Title Cultural Adaption	0.178	33.06	**24.28**	11.71	**13.51**
Llama3 Recipe Title Generated	0.132	20.21	13.65	7.03	9.01
Reranking					
Sentence-transformer + Llama3 Re-rank	0.193	20.47	18.66	10.27	**13.51**
CARROT (Rewriting + Re-ranking)					
CARROT-Llama3	**0.202**	**35.69**	22.93	**13.87**	**13.51**

Table 4: Evaluation on the cross-cultural recipe retrieval dataset with full lists (non-labelled samples are considered non-related), higher scores indicate better performance on all metrics. Please refer to Section 5.2 for details on the basic retrieval model, and for query rewrite and re-rank in section 4. **Bold** indicates the best performance across all method, underlined indicates best performance across all basic retrieval model.

- *Aubergine Parmigiana* and *Eggplant Parmesan* are **consistent**. Despite the difference in terminology, both names refer to the same dish.

Regarding the exact match with cultural adaptation, we allow greater differences in flavor and cooking steps, but these differences must reflect cultural variations.

The differences in recipes between different cultures are usually reflected in the following aspects:

- The selection of ingredients and seasonings will be more in line with the local culture
- The units for measuring ingredient quantities will differ
- The cooking methods and tools will be more suited to the local context.

For example:

- *Cucumber soup* can be interpreted differently across cuisines, in English recipes it could be cream-based cold soup, but in Chinese it could be hot soup with salty flavor. These differences reflect cultural variations
- *Chocolate drops* and *Chocolate cakes* have similar ingredients and flavor, but they can not be considered exact match because the differences can not reflect cultural variations .

Moreover, The results of an exactly matched recipe should not violate the user's explicit requirements regarding ingredients or flavors. For example:

- Source recipe is *Baby Food Cookies, No Salt, No Sugar Version* then results containing salt or sugar should not be considered an exact match.
- Source recipe's title is *Thai Green Curry* then a curry with Japanese flavors would not be an exact match.

Criteria of Partial Match Partially matched recipes are not fully similar to the source dish, but they are of referential value to the user and can provide some inspiration.

If two recipes have similar ingredients or flavors, and the differences between the two recipes do not exceed the scope that can provide referential value. they can be considered a partial match. The scope that can provide referential value refers to recipes belonging to the same category (for example, main courses, desserts, beverages, etc.).

- Although *Mapo Tofu(Spicy Tofu)* and *chili con carne* have different ingredients, their flavors are similar. Users can refer to the preparation process of spicy sauce when making chili pork sauce, therefore, they are considered a partial match.
- Although *chicken curry* and *Tuscan chicken stew* have different flavors, their main ingredients are consistent. They are considered partially related because other stewed chicken recipes can also provide certain references to users.

Criteria of Not Matching If two recipes neither meet the criteria for an exact match nor the criteria for a partial match, then they should be considered as not matching

For example, the differences between *rice pudding* and *streamed rice* are too significant to offer valuable references, so they are considered not matching each other.

B Prompt in CARROT Framework

B.1 Task A: Recipe Title Generation Task

```
Here is a Chinese recipe; please create a brief
English title for the recipe:
[Chinese recipe ingredients]
[Chinese recipe cooking steps]
```

B.2 Task B: Recipe Title Cultural Adaption Task

```
This is a Chinese recipe title, rewritten to
fit English cultural conventions:
[Chinese recipe title]
```

B.3 Task C: Recipe Re-ranking

```
  Given a Chinese recipe and some English
recipes, assess their relevance, and rank them
in the order of relevance. When the relevance
is the same, prioritize recipes that are more
aligned with the culture of English speakers.
[Relevance Instructions]: In Appendix A
[Chinese recipe]
[1][English recipe_1]
...
[n][English recipe_n]
(For Top1 Instruction): Select the identifier
of the most relevant English recipe
(Ranking Instruction): Listed the identifiers
in descending order of relevance
```

B.4 Task D: Generation Task

We follow the prompts in the previous work(Cao et al., 2024):

```
[Chinese recipe] Recipe in English, adapted to
an English-speaking audience:
```

C Model Version in the Experiment

We translate Chinese recipe to English by opus-mt models (Helsinki-NLP/opus-mt-zh-en), and retrieval English recipes by sentence-transformer (sentence-transformers/all-MPNet-base-v2) and we use colbert retrieval model (colbert-ir/colbertv2.0) and we also use multilingual sentence-transformer (distiluse-base-multilingual-cased-v1). We use bert-base-uncased(google-bert/bert-base-uncased) for calculating BertScore.

We explore various LLMs, include: Llama3-8B (meta-llama/Meta-Llama-3-8B-Instruct), Qwen1.5-7B (Qwen/Qwen1.5-7B-Chat), and Baichuan2-7B (baichuan-inc/Baichuan2-7B-Chat). All the models were run with default parameters.

D IR Results evaluation with full lists

Here we present the evaluation on the cross-cultural recipe retrieval dataset with full lists in Table 4. The conclusions of the table here are similar with results with condensed lists, shown in Section 6.3 and Table 2.

E Check List

Harmful information And Privacy We propose a Recipe Retrieval Dataset and we did not see any potential malicious or unintended harmful effects and uses, environmental impact, fairness considerations, privacy considerations, and security considerations in the work.

We also do not have data that contains personal information

License and Intend We provide the license we used here: Llama3(https://llama.meta.com/llama3/license/), Qwen1.5(https://huggingface.co/Qwen/Qwen1.5-7B-Chat/blob/main/LICENSE), Baichuan2 (Apache License 2.0), our use of these existing artifacts was consistent with their intended use.

Documentation of the artifacts We use the CulturalRecipes Dataset, it is in English and Chinese and annotated by six native Chinese speakers proficient in English with experience in both Chinese and Western cooking.