Cultural Adaptation of Recipes

Yong Cao[1,2*], **Yova Kementchedjhieva**[2*], **Ruixiang Cui**[2], **Antonia Karamolegkou**[2], **Li Zhou**[2,3], **Megan Dare**[4], **Lucia Donatelli**[4] **and Daniel Hershcovich**[2]

[1]Huazhong University of Science and Technology, China
[2]Department of Computer Science, University of Copenhagen, Denmark
[3]University of Electronic Science and Technology of China, China
[4]Department of Language Science and Technology, Saarland University, Denmark
{yongcao,yova,rc,antka,dh}@di.ku.dk, li_zhou@std.uestc.edu.cn,
{mdare,donatelli}@coli.uni-saarland.de

Abstract

Building upon the considerable advances in Large Language Models (LLMs), we are now equipped to address more sophisticated tasks demanding a nuanced understanding of cross-cultural contexts. A key example is recipe adaptation, which goes beyond simple translation to include a grasp of ingredients, culinary techniques, and dietary preferences specific to a given culture. We introduce a new task involving the translation and cultural adaptation of recipes between Chinese- and English-speaking cuisines. To support this investigation, we present CulturalRecipes, a unique dataset composed of automatically paired recipes written in Mandarin Chinese and English. This dataset is further enriched with a human-written and curated test set. In this intricate task of cross-cultural recipe adaptation, we evaluate the performance of various methods, including GPT-4 and other LLMs, traditional machine translation, and information retrieval techniques. Our comprehensive analysis includes both automatic and human evaluation metrics. While GPT-4 exhibits impressive abilities in adapting Chinese recipes into English, it still lags behind human expertise when translating English recipes into Chinese. This underscores the multifaceted nature of cultural adaptations. We anticipate that these insights will significantly contribute to future research on culturally aware language models and their practical application in culturally diverse contexts.

1 Introduction

Cooking recipes are a distinct form of procedural text whose accurate interpretation depends on several factors. Familiarity with ingredients and measurement units, common sense about the cooking environment, and reasoning about how tools and actions affect intermediate products in the cooking process are necessary to successfully craft a recipe. Such knowledge varies by culture and language, as a result of geography, history, climate, and economy (Albala, 2012). These factors impact the frequency of ingredient usage, the available forms and cost of heat for cooking, common taste profiles, written recipe style, etc. (§2).

Identifying and adapting to cultural differences in language use is important and challenging (Hershcovich et al., 2022). Recipe translations with current machine translation technology may gloss over culture-specific phraseology or yield mistranslations due to a lack of grounding in the physical and cultural space. Literal translations are often opaque or odd: a Chinese dish, 夫妻肺片 (fū qī fèi piàn) (literally, 'husband and wife lung slices'), can be adapted in translation to 'Sliced Beef in Chili Sauce' for English-speaking cooks. Structural patterns in recipes in different cultures (e.g., *mise en place*[1]) additionally make straightforward recipe translation difficult: cuisines differ in dish preparation methods, and temporal dependencies between actions complicate the disentanglement of recipe actions (Kiddon et al., 2015; Yamakata et al., 2017).

In this work, we introduce the task of adapting cooking recipes across languages and cultures. Beyond direct translation, this requires adaptation with respect to style, ingredients, measurement units, tools, techniques, and action order preferences. Focusing on recipes in Chinese and English, we automatically match pairs of recipes for the same dish drawn from two monolingual corpora,

*Equal contribution.

[1]In French cooking, *mise en place* is the practice of measuring out and cutting all ingredients in advance.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 80–99, 2024. https://doi.org/10.1162/tacl_a_00634
Action Editor: Taro Watanabe. Submission batch: 8/2023; Revision batch: 10/2023; Published 1/2024.

and train text generation models on these pairs. We evaluate our methodology with human judgments and a suite of automatic evaluations on a gold standard test set that we construct. We provide ample evidence that recipe adaptation amounts to more than mere translation and find that models finetuned on our dataset can generate grammatical, correct, and faithful recipes, appropriately adapted across cultures. Intriguingly, Large Language Models (LLMs) outperform our finetuned models in both automatic and human evaluations, even without training on our paired dataset. This unexpected result opens multiple avenues for future research, including how large-scale pre-training could complement our dataset and nuanced evaluation metrics that could better capture the complexities of recipe adaptation. Our contributions are as follows:

(a) We introduce the task of cross-cultural recipe adaptation and build a bidirectional Chinese-English dataset for it, **CulturalRecipes** (§3).

(b) We experiment with various sequence-to-sequence approaches to adapt the recipes, including machine translation models and multilingual language models (§6).

(c) We evaluate and analyze the differences between Chinese- and English-speaking cultures as reflected in the subcorpora (§4) and to the translation and adaptation of recipes (§6).

Our dataset, code, and trained models are available at `https://github.com/coastalcph/cultural-recipes`.

2 Cultural Differences in Recipes

Extensive cross-cultural culinary research reveals compelling differences in ingredients, measurement units, tools, and actions, each reflecting historical, geographical, and economic influences unique to each culture (Albala, 2012). For example, the historical reliance on open flame cooking in China has cultivated an array of oil-based cooking techniques exclusive to Chinese cuisine. Further complexities arise from culture-specific terminologies for cooking methods and dish names, which pose formidable challenges to translation and adaptation (Rebechi and da Silva, 2017). Additionally, the visual presentation of online recipes exhibits striking contrasts across different cultural contexts (Zhang et al., 2019a). Delving deeper, culinary preferences also demonstrate regional patterns in flavor profiles; Western cuisines tend to combine ingredients that share numerous flavor compounds, while East Asian cuisines often intentionally avoid such shared compounds (Ahn et al., 2011). These intricate cultural nuances underscore the complexity and diversity inherent in global culinary practices, thereby emphasizing the intricacy involved in adapting recipes across different cultures.

Figure 1: An example of cultural differences between Chinese (left) and English (right) recipes by color: blue text signals contrasts in ingredient measurement units; green, ingredients; orange, actions performed by cooks; and purple, tools. For readability, we show our literal translation on the left along with the original Chinese.

Examples. Figure 1 presents a Mandarin Chinese recipe and its human-authored adaptation to American English, highlighting key differences:

(1) *Ingredients.* Distinct ingredients feature prominently in each recipe; the Chinese version highlights 米酒 (mǐ jiǔ) 'rice wine', 红豆 (hóng dòu) 'red beans', and 带皮老姜 (dài pí lǎo jiāng) 'ginger with skin'. Interestingly, while 'red bean' is referenced in Chinese recipes, the equivalent ingredient is typically recognized as 'adzuki beans' in Western countries.

(2) *Measurement units.* Chinese recipes often rely on imprecise measurements, guided by the cook's experience, while American English recipes use precise U.S. customary or Imperial units like 'cups', 'inches', 'pints', and 'quarts'. Occasionally, Chinese recipes employ traditional units such as 两 (liǎng) and 斤 (jīn), or metric system units like 'grams (g)' and 'milliliters (mL)'.

(3) *Tools.* Specificity varies between recipes, with English recipes typically specifying pot sizes while Chinese recipes provide more general descriptions. Chinese recipes also favor stovetop cooking over ovens, contrasting with their English counterparts.

(4) *Actions by cook.* Preparation methods often vary between Chinese and English recipes. For instance, Chinese recipes usually involve shredding ginger, while English recipes recommend peeling and julienning. Additionally, unique processes like 焯水 'blanching', common in Chinese cooking to remove unwanted flavors, are rarely found in English recipes. These differences highlight the subtle cultural nuances in similar recipes.

Over-generalization and Bias. In a study of cultural adaptation, it is important to recognize that the concept of ''culture'' is multifaceted and complex. When we refer to Chinese- and English-speaking cultures throughout this work, we make the simplifying assumption that there are general features that characterize the cooking of these cultures and make them distinct in certain systematic ways. We recognize that there is enormous diversity within these simplistic categories,[2] but as a first step towards the adaptation of recipes across cultures, we restrict ourselves to the coarse-grained level only.

To enable the development and benchmarking of recipe adaptation, we build a dataset for the task.

3 The CulturalRecipes Dataset

Our dataset, *CulturalRecipes*, builds on two existing large-scale recipe corpora in English and Chinese, respectively. We create two collections of automatically paired recipes, one for each direction of adaptation (English→Chinese and Chinese→English), which we use for training and validation in our recipe adaptation experiments (§6). Additionally, *CulturalRecipes* incorporates a small test set of human adaptations expressly crafted for the task in each direction, serving as references in our experimental evaluation.

3.1 Recipe Corpora

We source recipes from two monolingual corpora: **RecipeNLG** (Bień et al., 2020) and **XiaChuFang** (Liu et al., 2022).[3] RecipeNLG consists of over 2M English cooking recipes. It is an extension of Recipe1M (Salvador et al., 2017) and Recipe1M+ (Marin et al., 2019), with improvements in data quality. XiaChuFang consists of 1.5M recipes from the Chinese recipe website `xiachufang.com`, split into a training and evaluation set. We use the training set and clean it by removing emojis,[4] special symbols, and empty fields. We use the title, ingredients, and cooking steps fields of the recipes from both corpora. The recipes in RecipeNLG consist of nine ingredients and seven steps on average, and in XiaChuFang, of seven ingredients and seven steps. As these two corpora are independent and monolingual, discovering recipe equivalents between them is not trivial.

3.2 Recipe Matching Rationale

Our recipe matching procedure relies on the following assumption: If two recipes have the same title, they describe the same dish. This assumption can be applied even in a monolingual context: if two recipes are both titled 'Veggie Lasagna', we can assume that they describe the same dish (Lin et al., 2020; Donatelli et al., 2021). It is permissible that there is some mismatch in the set of ingredients, in the number and sequence of steps, in the measurement units and exact amounts, etc. The same assumption can be said to hold for a recipe with a slightly different, but semantically equivalent title, e.g., 'Vegetable Lasagna'. Similarly, if we take the Chinese recipe title 卷心菜番茄牛肉汤, we translate it to 'Cabbage tomato beef soup' and we find a recipe with a very similar title in English, e.g., 'Cabbage beef soup', we can assume that these two recipes describe the same dish. The degree to which this assumption holds depends on the quality of translation of recipe titles from one language into the other, on the measure of similarity, and on how much distance we allow for between two recipe titles before they are no longer considered semantically equivalent. These factors guide our approach to building a silver-standard dataset for the task, further described below, with the procedure also

[2] For example, southern and northern Chinese cuisines are vastly different, with rice and wheat as staples, respectively.

[3] For license details, please refer to `https://recipenlg.cs.put.poznan.pl/dataset` for RecipeNLG and `https://xiachufang.com/principle` for XiaChuFang.

[4] Despite their potential significance, we remove emojis since they occur only in a few XiaChuFang recipes.

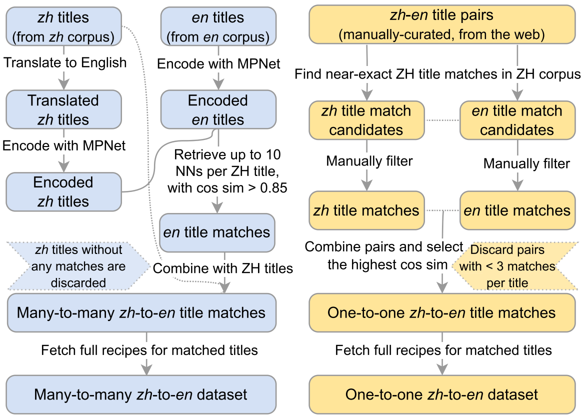

Figure 2: Training and validation (left) and test (right) silver-standard data compilation in the direction Chinese→English. The process is analogous for the opposite direction.

		# Recipes		Mean # Tokens	
		Source	**Target**	**Source**	**Target**
Train & Val	*zh*→*en*	44,5k	144,6k	159.1	140.2
	en→*zh*	43,8k	120,7k	117.1	164.8
Silver Test	*zh*→*en*	82	82	140.5	144.7
	en→*zh*	52	52	122.7	153.3
Gold Test	*zh*→*en*	25	25	139.8	97.1
	en→*zh*	41	41	115.7	176.5

Table 1: Statistics of (many-to-many) training, (one-to-one) silver-standard and gold-standard (human-written) evaluation sets for both directions. *zh*: Chinese. *en*: English. We count tokens with whitespace tokenization for English and `jieba` text segmentation for Chinese.

visualized in Figure 2, and the statistics of the resulting datasets reported in Table 1.[5]

3.3 Silver-standard Data

Training and Validation Sets. We obtain training recipe pairs by (1) automatically translating all recipe titles in the Chinese corpus to English using a pre-trained machine translation model (Tiedemann and Thottingal, 2020);[6] (2) encoding all English and translated Chinese titles with the MPNet sentence encoder (Song et al., 2020)[7] to obtain two embedding spaces; and (3) in each direction (English→Chinese and Chinese→English), retrieving up to $k = 10$ nearest neighbors per source title from the target space, and filtering out any neighbors that have a cosine similarity against the source title lower than 0.85.[8] The resulting sets, one in each direction, contain multiple reference targets for each source recipe. We further split the matches into training and validation sets.

We recognize that the aforementioned procedure can be susceptible to various sources of noise due to the translation of titles, the encoder representations, and the fixed similarity threshold. We trust that the signal-to-noise ratio should still be sufficient to enable model learning, but for evaluation we need cleaner, more representative data.

Test Set. We are able to eliminate one of the aforementioned sources of noise by collecting manual translations of Chinese recipe titles into English and vice versa from websites that explicitly mention the original dish name when presenting an adapted version.[9] This should resolve issues like 夫妻肺片 (fū qī fèi piàn) being translated literally by an automatic MT system (see §1). To supplement these titles with a corresponding list of ingredients and steps, we look up each title in the recipe corpus of the corresponding language and find the most similar title within, allowing for different capitalization, punctuation and slight differences in word choice and order, e.g., 'Rice with caramelized leeks' and 'Caramelized Leek Rice' (we manually inspect candidate matches to ensure semantic equivalence).

The resulting test set closely resembles the training data, thus allowing us to determine how well the models we train do in the setting they were trained for (mapping between automatically matched recipes). In order to evaluate the models' ability to perform the true task we want to solve, i.e. adapting specific recipes from one culture to another, we also construct a gold-standard test set.

[5] Prior to the procedure described below, we filter out recipes longer than 512 subword tokens (arbitrarily using the mT5 tokenizer; Xue et al., 2021) to facilitate using the neural approaches described in §6.

[6] `Helsinki-NLP/opus-mt-zh-en`.

[7] `sentence-transformers/all-mpnet-base-v2`.

[8] The similarity threshold for retrieval was chosen through manual inspection of the quality of retrieved pairs.

[9] For Chinese→English we use *Easy Chinese Recipes*, *Recipes Archives*, *Asian Food Archives*, *Authentic* Chinese Recipes; for English→Chinese, *Christine's Recipes* and Wikipedia. We convert any traditional Chinese text to simplified Chinese using `zhconv` to match our other data sources.

Figure 3: Screenshot from our human recipe adaptation platform, demonstrating the English→Chinese direction, with the source recipe on the left. On the right, participants should adapt the title, ingredients, and steps based on their culinary knowledge and cultural habits.

3.4 Gold-standard Test Data

We include human-written adaptations in our dataset as the ground truth for reference-based evaluations (§5.1, §5.2) and as a point of comparison in human evaluations (§5.3). We select 41 English recipes and 25 Chinese recipes manually from the silver test sets to adapt each to the other culture.

We develop an in-house web application as our recipe writing platform, illustrated in Figure 3. Our guidelines encourage participants to adapt recipes based on their culinary knowledge and cultural customs. We give participants the option to skip a recipe if they are not able to confidently adapt it. Six native Chinese speakers proficient in English with experience in both Chinese and Western cooking volunteered for the task, spending 6.4 minutes on average to adapt a recipe. Subsequently, three of the authors, fluent in both English and Chinese, who have substantial cooking experience, hand-corrected and improved all adapted recipes, including filtering incomplete source recipes, and correcting grammatical errors, spelling mistakes, and non-executable recipe expressions.

4 Corpus Analysis

Here, we perform a data-driven analysis to investigate how the cultural differences discussed in §2 are realized in English and Chinese recipe corpora through the lens of distributional semantics.

4.1 Embedding Alignment

In this analysis, we train static monolingual word embeddings on English and Chinese recipe data, respectively, as a means of capturing their distributional properties. While the global geometry of English and Chinese distributional spaces is similar (Lample et al., 2018), we hypothesize that cultural differences would lead to mismatches in the local geometry of the two spaces (Søgaard et al., 2018). We test this hypothesis through cross-lingual embedding alignment, wherein the English and Chinese embeddings are aligned through a linear mapping to obtain a cross-lingual embedding space, in which semantic equivalents between the two languages should occupy a similar position.

We train monolingual word embeddings using Word2Vec based on a skipgram model by Mikolov et al. (2013b) on the entire English and Chinese corpora (§3.3),[10] and align them using VecMap (Artetxe et al., 2017) with weak supervision from a seed dictionary of 15 culturally neutral word pairs we manually curate.[11]

4.2 Analysis

We use the top 100 most common Chinese content words in the XiaChuFang dataset (not included in our seed dictionary) as query terms and retrieve their five nearest neighbors in the English embedding space, thus inducing a bilingual lexicon from the cross-lingual embedding space (Mikolov et al., 2013a). We manually evaluate this dictionary for correct literal translations and report performance in terms of Precision@5: The ratio of query words for which the correct translation is among the word's five nearest neighbors in the target space (Lample et al., 2018). The equation is defined as:

$$\text{Precision@}k = \frac{N@k}{N}$$

where $N@k$ is the number of pairs with the correct literal translation in top k nearest neighbors and N is the total number of pairs.

The result is 68% (i.e., 68 of 100 query words were correctly mapped), which indicates that (a)

[10]We train 300-dimensional embeddings for 5 epochs using a minimum frequency count of 10, window size of 5, and 10 negative samples. Chinese text is tokenized with `jieba`.

[11]Seed dictionary: spinach-菠菜, onion-洋葱, flour-面粉, potatoes-土豆, egg-蛋, salt-盐, sugar-糖, apples-苹果, mix-混合, chop-劈, pour-倒, knife-刀, bowl-碗, pot-锅, chicken-鸡.

Source	Target	Nearest Neighbors
水果 (shuǐ guǒ)	fruit	fruit, fruits, kiwi, strawberry, seasonal
沙拉 (shā lā)	salad	feta, lebanese, bruschetta, tabbouleh, caesar
豆腐 (dòu fǔ)	tofu	boiled, ham, sausage, bacon, kielbasi
淀粉 (diàn fěn)	starch	flour, beaten, salt, shortening, pwdr
筷子 (kuài zi)	chopstick	fork, spatula, toothpick, wooden, knives
蒸 (zhēng)	steam	bake, 350, pans, boil, oblong

Table 2: Top-5 examples from bilingual lexicon induction with underlined literal matches, mismatches, and matches that can be attributed to cultural differences.

the global geometry of the two embedding spaces is indeed similar and VecMap has successfully aligned them using a seed lexicon of just 15 word pairs; and that (b) in the majority of the cases there is a 1:1 match between the Chinese and English words. More interesting, however, are the 32 words without a literal match. Here we find that 26 map onto what can be considered a cultural equivalent, while the other six can be considered accidental errors (due to lacking quality in the monolingual embeddings and/or inaccuracies in the alignment). We provide qualitative examples in Table 2.

A successful word match can be exemplified by 水果 (shuǐ guǒ) 'fruit', which correctly aligns with its English equivalent 'fruit' among the top five nearest neighbors. An instance of an inadvertent misalignment, however, can be observed with 沙拉 (shā lā) 'salad'. It is mapped closer to salad ingredients, other side dishes, and particular salad types, rather than precisely corresponding to the English term 'salad'.

Certain instances of misalignment can be attributed to cultural differences between English and Chinese culinary practices. Take for instance the ingredient 豆腐 (dòu fǔ) 'tofu', a staple protein source in Chinese cuisine, which aligns with 'ham', 'sausage', and 'bacon'—protein-rich food items prevalent in English-speaking cuisines. Similarly, 淀粉 (diàn fěn) 'starch' is matched with 'flour'. In terms of kitchen utensils, 筷子 (kuài zi) 'chopsticks' corresponds to 'fork', 'spatula', and 'toothpick', which perform comparable functions in Western culinary settings. Furthermore, the cooking technique 蒸 (zhēng) 'steam' maps onto 'bake', a heat-processing method more frequently used in English recipes. These examples underscore the cultural discrepancies between English and Chinese recipes, emphasizing that recipe adaptation goes beyond mere translation.

5 Cross-cultural Recipe Adaptation Task

We propose the task of cross-cultural recipe adaptation, which extends the task of machine translation with the requirement of divergence from the source text semantics in order to address cultural differences in the target culture. While translation studies have long considered culture (Bassnett, 2007), this is not yet explored in machine translation. Our matched cross-lingual corpora allow us to inform recipe adaptation by both language and culture simultaneously. In §6 we adopt an end-to-end sequence-to-sequence approach to the task to establish a set of baselines since this is the dominant approach in machine translation.

The evaluation of cultural adaptation should prioritize meaning preservation while allowing divergences in meaning as long as they stem from cross-cultural differences. This subjective criterion is challenging to implement, as cross-cultural differences, and by extension, the task itself, are not well-defined. As common in text generation tasks, we first adopt reference-based automatic evaluation metrics (§5.1). Furthermore, to capture structural similarity between references and predictions, we employ meaning representations for evaluation (§5.2). Crucially, since reference-based metrics are often unreliable for subjective tasks (Reiter, 2018), we additionally perform human evaluation (§5.3).

5.1 Surface-based Automatic Evaluation

We use various metrics to assess the similarity between the generated and reference recipes. We use three overlap-based metrics: BLEU (Papineni et al., 2002), a precision-oriented metric based on token n-gram overlap and commonly used in machine translation evaluation, ChrF (Popović, 2015), a character-level F-score metric that does not depend on tokenization,[12] and ROUGE-L (Lin, 2004), a recall-oriented metric based on longest common subsequences and widely used in summarization evaluation;[13] and one representation-based metric, BERTScore

[12]For BLEU and ChrF, we use `SacreBLEU` (Post, 2018) version 2.3.1 with default parameter settings.

[13]For evaluation, we replace newlines with spaces in all reference and generated recipes. We segment Chinese text to words with `jieba`.

(Zhang et al., 2019b), based on cosine similarity of contextualized token embeddings[14] and shown to correlate better with human judgments than the above metrics in various tasks.

5.2 Structure-aware Automatic Evaluation

Standard metrics may not effectively capture semantic similarity between texts due to sensitivity to surface form. To address this, we employ graph representations, a favored choice for capturing the flow of cooking actions, tool usage, and ingredient transformations in recipes (Mori et al., 2014; Kiddon et al., 2015; Jermsurawong and Habash, 2015; Yamakata et al., 2016). These allow for an examination of structural differences influenced by language and culture (Wein et al., 2022). Here, we leverage Abstract Meaning Representation (AMR; Banarescu et al., 2013), a general-purpose graph meaning representation, to represent recipes.

To generate AMR graphs, we employ `XAMR` (Cai et al., 2021),[15] a state-of-the-art cross-lingual AMR parser that can parse text from five different languages into their corresponding AMR graphs. It is based on a sequence-to-sequence model, utilizing mBART (Liu et al., 2020a) for both encoder and decoder initialization.

To assess the similarity between model-generated and reference texts' AMRs, we use the *Smatch* metric (Cai and Knight, 2013), which aligns both graphs and computes the F1 score that measures normalized triple overlap.

5.3 Human Evaluation

While the above automatic metrics provide quantifiable results, they inherently suffer from the limitation of depending on a fixed reference set. In reality, there exist multiple legitimate ways to adapt a recipe. To address this, we propose four criteria for human evaluation, which we conduct on the gold-standard test set.

We have evaluators assess the outputs from all methods, including the human-written adaptations, on four dimensions key to the cultural adaptation of recipes: (1) *Grammar*—The generated recipe is grammatically sound and fluent; (2) *Consistency*—The output aligns with the format of a fully executable recipe encompassing coherent title, ingredients, and cooking steps; (3) *Preservation*—The adapted recipe largely retains the essence of the source recipe, producing a dish akin to the original; (4) *Cultural Appropriateness*—The generated recipe integrates well with the target cooking culture, aligning with the evaluator's culinary knowledge and recipe style expectations. Evaluators mark each dimension on a 7-point Likert scale (Likert, 1932), where a higher score indicates superior performance. A single evaluator rates each recipe pair separately and independently.

Crowdsourcing Evaluation. We recruit evaluators on Prolific[16] and deploy our evaluation platform on the same in-house web application used for human recipe writing (§3.4). To ensure the evaluation validity, we require participants to be native speakers of the target language and proficient in the source language for each adaptation direction. Additionally, participants must successfully undergo a comprehension check, guided by our evaluation tutorial. Each evaluator is required to evaluate two example recipes for the comprehension check and three recipes for our tasks. This rigorous screening process secures the reliability and accuracy of the evaluations conducted for our study.

6 Experiments

Here we describe our recipe adaptation experiments and results, using the CulturalRecipes dataset introduced in §3. Due to their success in machine translation, we experiment with three end-to-end sequence-to-sequence classes of models to adapt recipes across cultures: (finetuned) machine translation models, finetuned multilingual encoder-decoder models, and prompt-based (zero-shot) multilingual language modeling. Additionally, we evaluate the automatic matching approach used in our dataset construction. These will serve as baselines for future work on this task.

6.1 Experimental Setup

We use our silver training set for finetuning in each direction and evaluate on both the silver and gold test sets. We represent a recipe as a concatenation of title, ingredients, and steps, each section prefixed with a heading

[14]We rely on `bert-base-uncased` for representing English text and `bert-base-chinese` for Chinese text.

[15]We use the trained AMR parser model from `https://github.com/jcyk/XAMR`.

[16]`https://www.prolific.co/`.

(‘`Title:`’, ‘`Ingredients:`’ and ‘`Steps:`’, for both English and Chinese recipes).[17]

Automatic Matching. Since the source recipes used in the creation of the gold-standard test set are a subsample of the ones found in the silver-standard test set, we have matches for them in the target language retrieved based on title similarity (see §3.3 for a reminder of how the silver-standard test set was constructed). We evaluate these retrieved matches against the gold-standard human-written references, to determine whether title-based retrieval is a viable method for recipe adaptation.

Machine Translation. Recognizing the intrinsic translation component of recipe adaptation between languages, we leverage pre-trained machine translation systems in our experiments. We experiment with `opus-mt` models (Tiedemann and Thottingal, 2020),[18] which show a strong performance in machine translation. We first evaluate them in zero-shot mode (MT-zs), that is, purely as machine translation models, and additionally after finetuning using our training and validation sets (MT-ft).

Multilingual Language Modeling. We finetune multilingual encoder-decoder pre-trained language models on the CulturalRecipes dataset. Such models perform well on translation tasks (Tang et al., 2020) and are generally trained on abundant monolingual as well as parallel data, so they could prove more suitable for the recipe domain and for our ultimate goal, recipe adaptation. We choose mT5-base (Xue et al., 2021),[19] a multilingual multitask text-to-text transformer pre-trained on a Common Crawl-based dataset containing 101 languages, and mBART50 (Tang et al., 2020),[20] a variant of mBART (Liu et al., 2020b) based on a multilingual autoencoder finetuned for machine translation.

Prompting LLMs. Building on the remarkable performance of Multilingual LLMs in zero-shot translation without additional finetuning or in-context learning (Wang et al., 2021), we explore their recipe translation and adaptation capabilities.

We use BLOOM (Scao et al., 2022), an LLM trained on the multilingual ROOTS corpus (Laurençon et al., 2022).[21] Using the ROOTS search tool (Piktus et al., 2023), we find it does not contain our recipe corpora. As BLOOM is an autoregressive language model trained to continue text, we prompt as follows for English→Chinese:

```
[English recipe] 中文菜谱，适合中国人的:
```

and for Chinese→English:

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

Further, we experiment with GPT-4 (OpenAI, 2023),[22] and ChatGLM2 (Zeng et al., 2022; Du et al., 2022),[23] state-of-the-art multilingual and Chinese instruction-tuned LLMs (Ouyang et al., 2022). While they have likely been trained on both our recipe corpora (§3.1), they do not benefit from our matching procedure (§3.3) or our newly written human-adapted recipes (§3.4). We prompt them as follows for English→Chinese:

```
Convert the provided English recipe into
a Chinese recipe so that it fits within
Chinese cooking culture, is consistent
with Chinese cooking knowledge, and meets
a Chinese recipe's style. [English recipe]
```

and for Chinese→English:

```
Convert the provided Chinese recipe into
an English recipe so that it fits within
Western cooking culture, is consistent
with Western cooking knowledge, and meets
a Western recipe's style. [Chinese recipe]
```

Technical Details. For finetuning, we use a batch size of 64 for MT-ft and 32 for mT5-base and mBART50; and a learning rate of 1e-4.[24] We set the maximum sequence length to 512 tokens and finetune models for 30 epochs with early stopping

[17] We treat these headings as language-invariant meta-text, which is removed in post-processing prior to evaluation.

[18] `Helsinki-NLP/opus-mt-{zh-en/en-zh}`.

[19] `google/mt5-base`.

[20] `facebook/mbart-large-50`.

[21] `bigscience/bloom-7b1`, a 7B-parameter model with a 2k-token length limit. Preliminary experiments showed poor results with BLOOMZ-7B, mT0-xxl-mt and FLAN-T5-xxl (Chung et al., 2022), which are finetuned on multitask multilingual prompts (Muennighoff et al., 2022)—they are biased towards short outputs, prevalent in their training tasks.

[22] `gpt-4-0314` via the OpenAI API (8k-token length limit).

[23] Accessed via FastChat (ChatGLM2-6B).

[24] Selected among the learning rates {1e-5, 1e-4} for MT-ft, {5e-5, 1e-4} for mT5-base and mBART50; and batch sizes {64, 128} for MT-ft and {32, 64} for mT5-base and mBART50.

Method	BLEU	ChrF	R-L	B-Sc	Smatch	# Tok.
			Chinese → English			
MT-zs	6.8	28.7	12.0	54.0	23.7	82.4
MT-ft	**68.9**	**43.8**	**22.3**	**64.6**	**33.1**	98.7
mT5	60.0	37.2	19.5	62.9	31.0	85.2
mBART50	44.5	36.0	21.0	63.4	32.1	89.9
			English → Chinese			
MT-zs	2.6	9.3	49.7	62.4	20.6	110.6
MT-ft	38.5	**37.1**	54.5	71.4	26.8	91.4
mT5	**39.2**	36.3	54.9	**71.9**	**27.0**	82.1
mBART50	30.5	32.9	**56.2**	71.1	25.5	103.2

Table 3: Automated evaluation results on the silver test sets using reference-based metrics: SacreBLEU (BLEU), ChrF, R-L (ROUGE), B-Sc (BERTScore)—all token-based, and *Smatch*—a structure-aware metric assessing AMR graph similarity. Higher scores indicate better performance on all metrics.

after 5 epochs of no improvement in BLEU on the silver validation set. We use two 40GB A100 GPUs for finetuning mT5 and mBART50 and a single one for finetuning MT-ft and for prompting BLOOM. We use the default settings for GPT-4. For ChatGLM2 we set the temperature to 0.7 and the maximum sequence length to 1024 tokens. For generation with all other models, we use a beam of size 3 and a repetition penalty of 1.2; we prevent repeated occurrences of any n-gram of length ≥ 5.

6.2 Results

Automatic Evaluation on the Silver Test Sets. As presented in Table 3, we restrict our evaluation on the silver-standard test set to finetuned methods,[25] as a sanity check for their quality under conditions resembling their training setting. We discern that finetuning the MT model considerably improves its performance across all metrics and in both adaptation directions. In Chinese→English, MT-zs emerges as the optimal foundation for finetuning, outperforming the other two methods, mT5, and mBART50, across all metrics. However, English→Chinese displays mixed outcomes, with diverse models excelling in different criteria. Structure-aware automatic evaluation results generally match other automatic results: MT-ft performs best on Chinese→English, while mT5-base performs best on English→Chinese.

[25]We include MT-zs as a reference point to observe the gains from finetuning this model to obtain MT-ft.

Method	BLEU	ChrF	R-L	B-Sc	Smatch	# Tok.
			Chinese → English			
MT-zs†	5.3	29.1	22.4	59.4	30.6	77.5
MT-ft	**28.0**	42.5	19.6	59.9	28.1	103.6
mT5	14.0	31.6	17.8	59.5	25.5	87.4
mBART50	10.2	33.9	19.7	60.5	27.3	93.2
BLOOM†	22.3	48.3	29.5	62.5	**33.7**	110.0
ChatGLM2	18.3	41.8	26.8	61.9	28.8	174.3
GPT-4†	**28.0**	**50.3**	**30.8**	**66.5**	33.4	216.6
Retrieval†	16.8	37.8	20.5	61.7	26.6	150.7
			English → Chinese			
MT-zs†	10.6	6.9	60.8	69.8	29.4	108.0
MT-ft	13.6	28.3	53.8	70.5	24.5	88.5
mT5	16.6	28.1	53.4	70.7	25.3	78.6
mBART50	11.8	25.4	54.8	69.7	23.5	100.3
BLOOM†	20.0	11.5	50.8	66.4	28.6	154.7
ChatGLM2	22.4	11.0	54.3	75.2	28.8	153.2
GPT-4†	21.1	21.9	**61.0**	**77.8**	**29.6**	213.3
Retrieval†	**32.8**	**33.6**	52.9	68.4	25.0	130.3

Table 4: Automatic reference-based evaluation results on the gold-standard human test sets. † indicates methods without training for the task (zero-shot).

Automatic Evaluation on the Gold Test Sets. Moving to the gold-standard test set results in Table 4, we gain further intriguing insights. The significant performance gap between MT-zs and MT-ft reemphasizes that the recipe pairs in our dataset are not merely translations of each other. Moreover, it underscores the systematic patterns in the matched pairs within our training corpus (reflecting the cultural adaptation of recipes) can indeed be learned via finetuning on retrieved recipes. In this scenario, the LLMs BLOOM, ChatGLM2, and GPT-4 outperform the finetuned methods. Particularly in the Chinese→English direction, LLMs consistently match or surpass the performance of the next best finetuned approach. Notably, a comparison of the average length of model predictions shows a tendency of LLMs to produce longer predictions than their counterparts, with GPT-4 generating double the number of tokens compared to other methods. Interestingly, the retrieval method scores are comparable to the finetuned models in both directions and sometimes even surpass them. Despite this, LLMs continue to prove more effective overall. *Smatch* scores show performance differences consistent with BERTScore across models for both silver and gold-standard test sets, with the exception that BLOOM slightly outperforms GPT-4 in Chinese→English.

Method	GRA	CON	PRE	CUL
Chinese → **English** (n = 25)				
MT-zs	2.6 ±1.5	2.4 ±1.7	2.3 ±1.4	2.7 ±1.6
MT-ft	4.5 ±1.8	3.7 ±2.0	3.0 ±2.1	4.3 ±2.1
mT5	4.1 ±2.1	3.8 ±2.1	3.2 ±2.2	3.7 ±2.2
BLOOM	3.3 ±2.0	3.3 ±2.0	3.4 ±2.0	2.8 ±1.8
ChatGLM2	4.1 ±2.4	4.3 ±2.2	4.6 ±2.1	4.0 ±2.3
GPT-4	**6.0** ±1.2	**6.1** ±1.3	**5.9** ±1.0	**6.0** ±1.2
Human	4.2 ±2.1	4.4 ±1.9	4.5 ±1.9	4.6 ±1.9
Retrieval	5.1 ±1.7	4.9 ±2.0	4.3 ±2.3	3.8 ±2.0
English → **Chinese** (n = 41)				
MT-zs	2.3 ±1.6	2.7 ±2.0	3.5 ±2.2	2.3 ±1.7
MT-ft	4.8 ±2.2	3.1 ±2.2	2.5 ±1.9	3.2 ±2.0
mT5	4.3 ±2.0	3.4 ±2.1	2.8 ±2.0	3.5 ±1.9
BLOOM	3.8 ±2.1	4.2 ±2.1	4.6 ±1.9	3.0 ±1.6
ChatGLM2	5.4 ±1.7	**5.3** ±1.7	**5.7** ±1.6	4.1 ±2.3
GPT-4	5.3 ±2.0	5.1 ±2.0	5.2 ±1.9	**4.4** ±2.0
Human	**5.8** ±1.1	5.1 ±1.9	5.5 ±1.6	4.3 ±1.8
Retrieval	4.5 ±1.9	3.9 ±2.0	3.3 ±2.0	3.5 ±1.7

Table 5: Human evaluation results on the gold-standard test sets: average and standard deviation across recipes for each method and metric, ranging from 1 to 7. Note that different participants manually adapted (''Human'') and evaluated the recipes.

Human Evaluation. Table 5 showcases the results of human evaluation, with abbreviations GRA, CON, PRE, and CUL representing Grammar, Consistency, Preservation, and Cultural Appropriateness, respectively.[26] GPT-4 excels significantly across all metrics in the Chinese→English direction, even surpassing explicit human adaptation. Recipes retrieved from popular websites are a close second in GRA and CON, reflecting their high quality. However, the targeted adaptations written by humans who were explicitly instructed to adapt the source recipe to the target culture, perform better in PRE and CUL. For English→Chinese, GPT-4 remains the top performer only in CUL, while mT5 parallels the retrieved recipes in this metric. Notably, ChatGLM2 surpasses even human writers in CON and PRE, but not in GRA.

Correlation of Automatic Metrics with Humans. To determine the reliability of automatic metrics in assessing the quality of recipe adaptations, we examine their correlation with human

[26]We exclude mBART50 due to its architectural and performance similarity to mT5.

	BLEU	ChrF	R-L	B-Sc	Smatch
Chinese → **English**					
GRA	0.135	0.250*	0.135	0.257*	0.021
COR	0.151	0.268*	0.180	0.294*	0.065
PRE	0.174	0.312*	0.261*	0.260*	0.176
CUL	0.120	0.216*	0.189	0.237*	0.071
avg.	0.153	0.255*	0.202*	0.277*	0.079
English → **Chinese**					
GRA	0.286*	0.353*	0.201*	0.278*	0.070
COR	0.227*	0.232*	0.183*	0.217*	0.116
PRE	0.268*	0.180*	0.218*	0.247*	0.124
CUL	0.216*	0.268*	0.155	0.219*	0.081
avg.	0.290*	0.295*	0.221*	0.272*	0.117

Table 6: Kendall correlation of human evaluation results with automatic metrics. Statistically significant correlations are marked with *, with a confidence level of $\alpha = 0.05$ before adjusting for multiple comparisons using the Bonferroni correction (Bonferroni, 1936).

evaluations across the four metrics and their average. We use Kendall correlation, which is the official meta-evaluation metric used by WMT22 metric shared task (Freitag et al., 2022).

As illustrated in Table 6, all cases exhibit a positive correlation, albeit with varying strengths from weak to moderate, and with inconsistent performance between the two adaptation directions. For Chinese→English, ChrF and BERTScore indicate the strongest correlation with the average of all criteria. BERTScore further stands out by demonstrating the highest correlation with each individual criterion. On the other hand, for English→Chinese, BLEU performs comparably well, thus highlighting that the effectiveness of these metrics can vary based on the direction of adaptation. ROUGE-L, however, displays a significantly lower correlation, suggesting its limitations in evaluating recipe adaptations. Finally, we observe that *Smatch* is not significantly correlated with human judgments, possibly due to noise introduced by parsing errors.[27]

CUL presents the weakest correlation with most automatic metrics, underscoring the current limitations of automated evaluations in assessing the

[27]Inspecting XAMR outputs, we notice recurrent errors in both languages, likely attributable to the unique recipe genre. Common culinary actions are often incorrectly represented or overlooked: in English, actions like 'oil' or 'grease' are treated as objects. Similarly in Chinese, many actions are often omitted or associated with unrelated concepts.

cultural alignment of recipes, and highlighting the essential role of human evaluators. Notably, correlations for English→Chinese generally exhibit greater strength than Chinese→English. This discrepancy is likely due to the variation in sample sizes between the two directions.

7 Analysis and Discussion

Our findings reinforce previous research asserting the cultural bias of LLMs—specifically GPT-4—towards Western, English-speaking, U.S. culture, as exemplified in the food domain (Cao et al., 2023; Naous et al., 2023; Keleg and Magdy, 2023; Palta and Rudinger, 2023). However, our results also offer a more nuanced perspective. While GPT-4 demonstrates an exceptional ability to adapt to Chinese cuisine, its linguistic and semantic capabilities are outperformed by ChatGLM2 in English→Chinese. To delve deeper into these intriguing results, this section examines the strategies these models employ in the adaptation task.

Quantitative Analysis. Referring back to the analysis from §4, we choose a subset of six words and examine how they are handled by four models (MT-zs, MT-ft, and mT5, and GPT-4). Specifically, we measure the rate of literal translation of these concepts by each model, in the context of the recipes from the silver-standard test set of CulturalRecipes.[28] For instance, in adapting from English to Chinese, we identify *baking* as an English-specific concept. We count the appearances of related terms such as ‘bake’, ‘roast’, ‘broil’, and ‘oven’ in English source recipes, denoted as c_{source}. For each instance, we tally the occurrences of the direct translation, 烤 (kǎo), in the corresponding Chinese recipes, denoted as c_{target}, from either model predictions or retrieved references. We calculate the literal translation rate as $\frac{c_{target}}{c_{source}}$. Figure 4 visualizes the results for five culturally specific concepts and a universally applicable concept, ‘oil’.

We include ‘oil’ as a sanity check and indeed see that the literal rate of translation is high in both the references and in all model predictions.

The references show a low to medium rate of literal translations for the remaining five concepts, confirming their cultural specificity. MT-zs often

[28]We use the silver-standard test set rather than the gold-standard test set for its comparatively larger size.

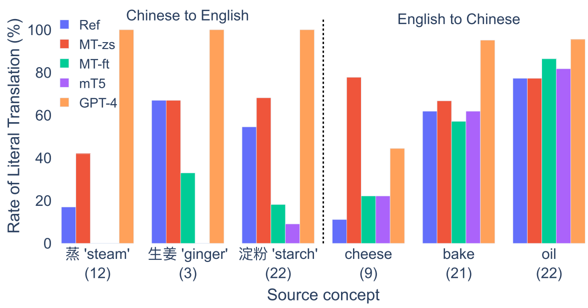

Figure 4: Analysis of the translation of specific concepts by the different models on the silver-standard test data. Ref = retrieved reference. In brackets, we show the number of occurrences of each concept.

translates these concepts literally, as could be expected from a machine translation model designed for near-literal translation—the difference is especially noticeable for the concepts ‘steam’ and ‘cheese’. The finetuned models MT-ft and mT5, on the other hand, learn to avoid literal translation, presumably opting for culturally appropriate alternatives instead—for ‘steam’, for example, none of the 12 occurrences of the concept in the source Chinese recipe are literally translated in the predictions of MT-ft and mT5.

An interesting trend emerges in GPT-4 predictions, where literal translations are found at a high rate for all concepts, often close to 100%. While this seems counter-intuitive considering the goal of adapting the culturally specific ingredients and cooking methods, in the next section we find that GPT-4 employs a slightly different strategy than just substituting these ingredients and methods.

Qualitative Analysis. We present a qualitative analysis highlighting the adaptation strategies adopted by models, specifically MT-zs, MT-ft, and GPT-4. The analysis centers on the Chinese recipe shown in Figure 1, with model predictions shown in Table 7. The translation from **MT-zs** directly incorporates Chinese ingredients not common in English recipes, accompanied by numerous spelling and grammatical errors. The prevalence of errors can be attributed to a dearth of recipe domain representations in the machine translation training data of MT-zs. In contrast, **MT-ft** offers a notably improved recipe rendition, albeit a wholly different red bean soup from the source recipe. Although this results in minimal content retention, it can be viewed as an extreme cultural adaptation, given the infrequent appearance

MT-zs: Title: Red beans soup, Ingridients: red beans, rice wine, leather ginger, Steps: Chrysotile, and red beans in the wine, soaked for eight hours, soaked in the pot, boiled in fire, stired in fire, turned into small fires for 30 minutes.
MT-ft: Title: Red Bean Soup, Ingredients: 2 tablespoons olive oil 1 large onion, chopped 2 garlic cloves, minced 1 teaspoon dried oregano 1 teaspoons dried basil 1 teaspooon dried thyme 1/2 teaspoon salt 1/4 teaspoon pepper 2 cups red kidney beans, rinsed and drained 1 (14 1/2 ounce) can diced tomatoes, undrained 1 (15 ounce) can red kidney beans, un-drained 1 (14 ounce) can beef broth 1 cup water 1/4 cup chopped fresh parsley, Steps: Heat oil in a large saucepan over medium heat. Add onion and garlic; cook and stir until tender. Stir in oregano, basil, thyme, salt and pepper. Add beans, tomatoes, broth, water and parsley; bring to a boil. Reduce heat; cover and simmer for 30 minutes or until beans are tender. Remove from heat; let stand for 5 minutes. Stir in parsley.
GPT-4: Title: Red Bean Soup Ingredients: - 1 cup red beans (adzuki beans) - 1/4 cup rice wine (or substitute with dry sherry) - 1-inch piece of fresh ginger, with skin Instructions: 1. Rinse the red beans and place them in a bowl. Thinly slice the ginger and add it to the bowl with the beans. 2. Pour the rice wine over the beans and ginger, ensuring they are fully submerged. Stir to combine, then cover and let soak for 8 hours or overnight. 3. After soaking, transfer the beans, ginger, and soaking liquid to a large pot. Bring the mixture to a boil over high heat, stirring occasionally. 4. Once boiling, reduce the heat to low and let the soup simmer for 30 minutes, stirring occasionally to prevent sticking. 5. Taste the soup and adjust the seasoning if necessary. Serve hot and enjoy your Red Bean Soup!

Table 7: Case study: English adaptations of the Chinese recipe from Figure 1, with manually highlighted (spelling, grammar or semantic) errors, adaptations to cultural differences, and failures to account for such.

of sweet red bean soup in Western cuisine. However, MT-ft sporadically manifests consistency errors, exemplified in this case by duplicating beans in the ingredient list and parsley in the steps. These findings confirm that the generation of coherent recipes continues to be a challenging endeavor for sequence-to-sequence models, corroborating the findings of prior work (Li et al., 2022).[29] **GPT-4**, on the other hand, generates a recipe more closely aligned with the source than the human-generated reference (refer to Figure 1). This model also incorporates thoughtful cultural adaptations: It quantifies ingredient amounts, unlike the source which vaguely indicates ''适量'' (shì liàng) (*moderate amount*), and it provides alternative names or substitutions for uniquely Chinese ingredients. The recipe instructions retain the crucial details from the source recipe, whilst maintaining fluency and appropriateness for Western-style recipes.

8 Related Work

Cultural Adaptation of Text. Cultural adaptation overlaps with style transfer, where the goal is to change the style of text while preserving the meaning (Jin et al., 2022). In addition to style, cultural adaptation also concerns common ground, values and topics of interest (Hershcovich et al., 2022). Particularly in culture-loaded tasks, it becomes crucial to consider cultural differences (Zhou et al., 2023a,b). While semantic divergences are usually treated as errors in machine translation (Briakou and Carpuat, 2021), cross-cultural translation often requires adaptations that change the meaning, e.g., by adapting entities (Peskov et al., 2021) or by adding explanations (Kementchedjhieva et al., 2020). We share the motivation of this line of work, but for the first time focus on recipes, where cultural adaptation is grounded in clear goals (accessibility to the cook and quality of the resulting dish).

Recipe Generation. van Erp et al. (2021) outline potential cross-disciplinary approaches involving NLP and food science, claiming that the analysis of digital recipes is a promising but challenging task. Marin et al. (2019) introduce the Recipe1M dataset (see §3) and Lee et al. (2020) finetune GPT-2 (Radford et al., 2019) on it to create a large English language model, RecipeGPT, capable of generating cooking instructions from titles and ingredients or ingredients from instructions and titles. Majumder et al. (2019) introduce a dataset of 180K English recipes from the website Food.com and a neural model to generate recipes according to user preferences inferred from historical interactions. Contrary to these, we focus on recipe adaptation, where generation is conditioned on a source recipe.

Recipe Adaptation. Donatelli et al. (2021) align recipes for the same dish on the action level using recipe graphs (Yamakata et al., 2016), aiming to adapt recipes to users of different levels of expertise. Morales-Garzón et al. (2021a,b, 2022) propose an unsupervised method to adapt recipes according to dietary preferences by proposing ingredient substitutions using domain-specific word and sentence embeddings. However, they do not modify the recipe steps beyond simple ingredient substitution. Li et al. (2022) build a dataset of 83K automatically-matched recipe pairs for the task of editing recipes to satisfy dietary restrictions. They train a supervised model to perform controlled generation, outperforming RecipeGPT.

[29] Similar behavior is observed in the other sequence-to-sequence models trained on our training set and in the automatically matched (retrieved) recipe.

They identify the remaining challenge of ''controllable recipe editing using more subtle traits such as cuisines (e.g., making a Chinese version of meatloaf)'', which we address here. Antognini et al. (2023), in contrast, propose addressing the same task *without* paired data, utilizing an unsupervised critiquing module and also outperforming RecipeGPT in both automatic and human evaluation. Liu et al. (2022) present a dataset of 1.5M Chinese recipes and evaluate compositional generalization in neural models in the task of counterfactual generation of recipes with substituted ingredients. They find recipe adaptation to be a challenging task: language models often generate incoherent recipes or fail to satisfy the stated constraints. In contrast, we find that after finetuning pre-trained models on our dataset, the models succeed in the task of cultural adaptation.

9 Conclusion and Future Work

In this work, we studied the task of adapting cooking recipes across cultures. We identified dimensions relevant to this task through a data-driven analysis, including differences in ingredients, tools, methods, and measurement units. We introduced CulturalRecipes, a dataset of paired Chinese and English recipes, and evaluated various adaptation methods. Through our experiments and analysis, we show that models can learn to consider cultural aspects, including style, when adapting recipes across cultures, with some challenges remaining in the level of detail and consistency between the different components of a recipe.

We envision our dataset and baselines will be useful for both downstream applications and further studies of cultural adaptation within and beyond NLP. Automatically adapting recipes from one culture to another could facilitate cross-cultural cross-pollination and broaden the horizons of potential users, serving as a bridge between people through food, and being useful to both novice and experienced cooks. Furthermore, our dataset is a challenging benchmark for language models: Besides the complex compositional generalization ability required for recipe adaptation (Liu et al., 2022), it assesses the ability of multilingual language models to adapt to target cultural characteristics, and to construct well-formed and faithful recipes. Lastly, our cross-cultural comparative analysis can be extended to sociological and anthropological research.

Future Work. As acknowledged in §2, the cultural categories we assume are highly simplistic. Future work will expand our datasets to treat finer-grained differences, as well as broaden it to more languages and cultures. It will further investigate the factors that impact recipe adaptation and develop more sophisticated modeling approaches to consider them, beyond the sequence-to-sequence approaches we experimented with here. Finally, our dataset can provide a starting point for related tasks, including recipe classification and retrieval.

Cultural categorization can be a sensitive topic so we have been careful to approach it with respect for the communities involved; we encourage future research in the area to maintain this practice. We hope that our research can contribute to a greater understanding and appreciation of diverse cultural traditions and practices related to food and cooking.

Acknowledgments

The authors extend their sincere gratitude to the reviewers and action editors for their invaluable feedback, which significantly contributed to the improvement of this work. Special thanks are also due to Laura Cabello and Nicolas Garneau for their insightful comments and to Qinghua Zhao and Jingcun Huang for their valuable assistance during our initial human evaluations. We also extend our sincere appreciation to the Department of Food Science, University of Copenhagen, and especially Qian Janice Wang for their contributions as volunteer human review adapters. The authors gratefully acknowledge the HPC RIVR consortium (www.hpc-rivr.si) and EuroHPC JU (eurohpc-ju.europa.eu) for funding this research by providing computing resources of the HPC system Vega at the Institute of Information Science (www.izum.si). Yong Cao and Li Zhou gratefully acknowledge financial support from China Scholarship Council. (CSC No. 202206070002 and No. 202206160052).

References

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, and Albert-László Barabási. 2011.

Flavor network and the principles of food pairing. *Scientific Reports*, 1(1):196. https://doi.org/10.1038/srep00196, PubMed: 22355711

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

Diego Antognini, Shuyang Li, Boi Faltings, and Julian McAuley. 2023. Assistive recipe editing through critiquing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 375–384, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.28

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462. https://doi.org/10.18653/v1/P17-1042

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Susan Bassnett. 2007. Culture and translation. *A Companion to Translation Studies*, pages 13–23. https://doi.org/10.21832/9781853599583-003

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. https://doi.org/10.18653/v1/2020.inlg-1.4

Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7236–7249, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.562

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2021. Multilingual AMR parsing with noisy knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2778–2789, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.237

Shu Cai and Kevin Knight. 2013. Smatch: An evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.

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. https://doi.org/10.18653/v1/2023.c3nlp-1.7

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Lucia Donatelli, Theresa Schmidt, Debanjali Biswas, Arne Köhn, Fangzhou Zhai, and Alexander Koller. 2021. Aligning actions across recipe graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6930–6942, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.554

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Marieke van Erp, Christian Reynolds, Diana Maynard, Alain Starke, Rebeca Ibáñez Martín, Frederic Andres, Maria C. A. Leite, Damien Alvarez de Toledo, Ximena Schmidt Rivera, Christoph Trattner, Steven Brewer, Carla Adriano Martins, Alana Kluczkovski, Angelina Frankowska, Sarah Bridle, Renata Bertazzi Levy, Fernanda Rauber, Jacqueline Tereza da Silva, and Ulbe Bosma. 2021. Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food. *Frontiers in Artificial Intelligence*, 3. https://doi.org/10.3389/frai.2020.621577, PubMed: 33733227

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). 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. https://doi.org/10.18653/v1/2022.acl-long.482

Jermsak Jermsurawong and Nizar Habash. 2015. Predicting the structure of cooking recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 781–786, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1090

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205. https://doi.org/10.1162/coli_a_00426

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.389

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. https://doi.org/10.18653/v1/2020.coling-main.180

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 982–992. https://doi.org/10.18653/v1/D15-1114

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Helena H. Lee, Ke Shu, Palakorn Achananuparp, Philips Kokoh Prasetyo, Yue Liu, Ee-Peng Lim, and Lav R. Varshney. 2020. RecipeGPT: Generative pre-training based cooking recipe generation and evaluation system. In *Companion Proceedings of the Web Conference 2020*, pages 181–184. `https://doi.org/10.1145/3366424.3383536`

Shuyang Li, Yufei Li, Jianmo Ni, and Julian McAuley. 2022. SHARE: A system for hierarchical assistive recipe editing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11077–11090, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.emnlp-main.761`

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.

Angela Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepta Dey, and Bill Dolan. 2020. A recipe for creating multimodal aligned datasets for sequential tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4871–4884, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.440`

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. `https://doi.org/10.18653/v1/2022.emnlp-main.497`

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. `https://doi.org/10.1162/tacl_a_00343`

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. `https://doi.org/10.1162/tacl_a_00343`

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. Generating personalized recipes from historical user preferences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5976–5982, Hong Kong, China. Association for Computational Linguistics. `https://doi.org/10.18653/v1/D19-1613`

Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. `https://doi.org/10.48550/ARXIV.1309.4168`

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Neural and Information Processing System (NIPS)*.

Andrea Morales-Garzón, Juan Gómez-Romero, and Maria J. Martin-Bautista. 2021a. Semantic-aware transformation of short texts using word embeddings: An application in the food computing domain. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 148–154, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-srw.20

Andrea Morales-Garzón, Juan Gómez-Romero, and Maria J. Martín-Bautista. 2022. Contextual sentence embeddings for obtaining food recipe versions. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 306–316, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-031-08974-9_24

Andrea Morales-Garzón, J. Gómez-Romero, and M. J. Martin-Bautista. 2021b. A word embedding-based method for unsupervised adaptation of cooking recipes. *IEEE Access*, pages 1–1. https://doi.org/10.1109/ACCESS.2021.3058559

Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. 2014. Flow graph corpus from recipe texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2370–2377, Reykjavik, Iceland. European Language Resources Association (ELRA).

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*. https://doi.org/10.18653/v1/2023.acl-long.891

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

OpenAI. 2023. GPT-4 technical report. http://arxiv.org/abs/2303.08774

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, Sabu John, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.631

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. https://doi.org/10.3115/1073083.1073135

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. https://doi.org/10.18653/v1/2021.findings-emnlp.315

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The ROOTS search tool: Data transparency for LLMs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 304–314, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-demo.29

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. https://doi.org/10.18653/v1/W15-3049

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rozane Rodrigues Rebechi and Márcia Moura da Silva. 2017. Brazilian recipes in Portuguese and English: The role of phraseology for translation. In *Computational and Corpus-Based Phraseology*, pages 102–114, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-69805-2_8

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401. https://doi.org/10.1162/coli_a_00322

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3020–3028. https://doi.org/10.1109/CVPR.2017.327

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pámies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Nikolaus Muellner, Pascale Fung, Patrick Haller, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics. `https://doi.org/10.18653/v1/P18-1072`

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. https://doi.org/10.48550/ARXIV.2008.00401

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT—Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Wenxuan Wang, Maosong Sun, and Yang Liu. 2021. Language models are good translators. https://doi.org/10.48550/ARXIV.2106.13627

Shira Wein, Wai Ching Leung, Yifu Mu, and Nathan Schneider. 2022. Effect of source language on AMR structure. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 97–102, Marseille, France. European Language Resources Association.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

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. https://doi.org/10.1145/3106668.3106672

Yoko Yamakata, Shinji Imahori, Hirokuni Maeta, and Shinsuke Mori. 2016. A method forextracting major workflow composed of ingredients, tools, and actions from cooking procedural text. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. https://doi.org/10.1109/ICMEW.2016.7574705

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Qing Zhang, Christoph Trattner, Bernd Ludwig, and David Elsweiler. 2019a. Understanding cross-cultural visual food tastes with online recipe platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 671–674. https://doi.org/10.1609/icwsm.v13i01.3270

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019b. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023a. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.c3nlp-1.2

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023b. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. *arXiv preprint arXiv:2310.06458*. https://doi.org/10.18653/v1/2023.c3nlp-1.2