Do LLMs Understand Wine Descriptors Across Cultures? A Benchmark for Cultural Adaptions of Wine Reviews

Anonymous ACL submission

Abstract

Leveraging remarkable advancements in Large Language Models (LLMs), we are now poised to tackle increasingly complex challenges requiring deep comprehension of multifaceted domains and contexts. A specific application scenario is wine reviews adaptation. Wine reviews usually describe a wine's appearance, aroma, and flavor to help consumers appreciate its characteristics. However, the adaptation of wine reviews transcends mere translation; it requires consideration of regional preferences, flavor descriptors, and cultural nuances that shape wine perception. We introduce the first-ever task involving the translation and cultural adaptation of wine reviews between Chinese and English. In a case study on cross-cultural wine review adaptation, we compile the **first** dataset of 8k Chinese and 16k Western professional wine reviews. We evaluated various methods, including LLMs and traditional machine translation techniques, using both automatic and human metrics. For human assessments, we introduce three novel culture-related metrics—Cultural Proximity, Cultural Neutrality, and Cultural Genuineness—to gauge the success of different approaches in achieving authentic cross-cultural adaptation. Our analysis shows that current models struggle to capture cultural nuances, especially in translating wine descriptions across different cultures. This highlights the challenges and limitations of translation models in handling cultural content.

1 Introduction

Wine reviews serve as valuable guides for consumers, offering detailed insights into the characteristics of each bottle. For casual drinkers and connoisseurs alike, these reviews act as a compass, helping them navigate the vast selection of wines available. However, due to cultural influences and geographical distinctions, beverage consumption patterns and individual preferences vary significantly across regions (Rodrigues and Parr, 2019),

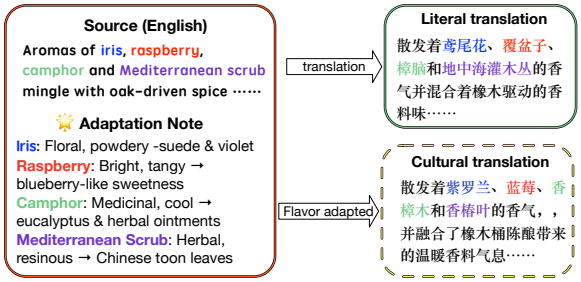


Figure 1: Clture-specific flavor misinterpretation in translation. Color coding highlights shifts in meaning: blue represents floral, red represents fruit, green represents spice, and purple represents vegetal reinterpretations. The figure shows how culturally unfamiliar terms may be recontextualized in target-language adaptations.

not to mention reviews. These variations extend beyond mere differences in taste and are deeply rooted in the cultural, social, and environmental contexts of each region. Consequently, consumer preferences for certain beverages, including wine, can differ greatly, often rendering generalized reviews less relevant or applicable across diverse audiences. Professional reviews play a greater role than user reviews in promoting consumer purchases (Chiou et al., 2014). For Chinese wine consumers, professional wine reviews in Chinese are scarce, and most available reviews require a paid subscription. Similarly, Western consumers face challenges finding professional reviews of Chinese wines. Yet, interest in Chinese wine is rapidly growing (Wine Intelligence, 2024; China, 2023). Consumer-research evidence underscores why accessible, comprehensible reviews matter (Danner et al., 2017; Spence, 2020). These trends underscore the increasing global relevance of Chinese wine and highlight the need for better information accessibility and cross-cultural understanding. Our initial goal is to address the following research question **RQ1:** To what extent do Chinese and Western professional wine reviews differ? To an-

swer **RQ1**, we propose the first cross-cultural multilingual wine reviews dataset CulturalWR, covering over 5k wines and about 25k reviews. We then perform comprehensive data analysis on it to analyze lexical and semantic differences. We demonstrate that Chinese and Western reviewers tend to emphasize different flavor descriptor and adopt divergent stylistic conventions in professional writing. These findings highlight the need for cultural adaptation when translating wine reviews across languages.

Recognizing and adapting to cultural differences in language use is both essential and challenging (Hershcovich et al., 2022), especially for subjective comments. Translations of reviews using current neural machine translation systems may overlook culture-specific expressions or result in mistranslations due to insufficient grounding in physical and cultural contexts. For instance, in Figure 1, 'raspberry', a common European flavor descriptor, is rare in China ('覆盆子'), making its taste unfamiliar to Chinese consumers. The flavor of raspberry is puzzling to Chinese consumers, and requires additional explanation or finding a fruit with a similar flavor. 'Blueberry' which has a similar flavor profile (Jin et al., 2022b), is more popular and better understood by Chinese consumers. All the outputs of this example are shown in Appendix I.

Based on the findings, we are interested in answering the research question **RQ2:**, To what extent do cross-cultural flavor-descriptor differences hinder target-culture consumers' comprehension of wine reviews? To measure this, we introduce three cultural-related evaluation metrics - Cultural Proximity, Cultural Neutrality and Cultural Genuineness and conduct human assessment to analyze different cultures' reviews from these perspectives. We found that readers most easily understand reviews rooted in their own culture. Surprisingly, Western reviews are harder for Chinese readers, whereas Chinese reviews are rated slightly easier to understand by Western readers.

Finally, we are interested in answering the research question **RQ3:** To what extent do LLM-generated culturally adapted wine reviews enhance cross-cultural alignment and reader comprehension compared to human translations? To answer this question, we prompted the LLM with culture-aware instructions to translate the reviews and then conducted a blind human evaluation. According to the annotators' ratings, the LLM can, under appropriate prompting, sometimes outperform literal human translations—a surprising outcome.

In this work, we introduce the task of adapting wine reviews across languages and cultures. Beyond direct translation, this requires adaptation concerning content and style. We focus on Chinese and English wine reviews, automatically pairing reviews for the same wine from the same vintage from two monolingual corpora. As there are many reviews in English for the same wine, we also explore the inner difference in reviews from people of the same culture. We evaluate our methodology with human evaluation and automatic evaluations on the dataset we construct. Our contributions are as follow:

1. We introduce the novel task of cross-cultural wine review translation and build the **first** bidirectional Chinese-English dataset with multiple references for it: CulturalWR.

2. We conduct a detailed evaluation and analysis of the cultural differences between Chinese and English-speaking communities in the lexical choices, semantic difference, and sensory descriptions used in wine reviews.

3. We experiment with various sequence-to-sequence approaches to adapt the reviews, including machine translation models and multilingual LLMs.[1]

2 Related Work

Computational analysis of wine reviews. Recent advancements in wine informatics have leveraged computational techniques to analyze expert wine reviews. The Computational Wine Wheel 2.0 facilitates machine-learning-based wine-attribute analysis (Chen et al., 2016). Machine-learning and text-mining techniques have been utilized to discover predictive patterns in wine descriptions, challenging the notion that flavor descriptions are purely subjective (Lefever et al., 2018). Recent work has begun to interrogate linguistic and cultural variation in wine discourse. (Hörberg et al., 2025) used a 68-million-word corpus and transformer language models to map chemosensory vocabularies across wine, perfume and food reviews, revealing domain-specific semantic clusters. Cross-lingual resources have also emerged: Wang et al. (2021) released an English–Chinese parallel corpus of 1,211 aligned reviews, enabling systematic

[1]Data and code available in https://github.com/anonymous-submission-papercode/CulturalWR

comparison of culture-specific tasting terms (Wang et al., 2021). Sensory studies confirm that cultural familiarity shapes perceived quality, as shown in a comparative experiment with British and Spanish experts (Suárez et al., 2023). These studies provide a foundation for computational analysis of wine reviews, yet they primarily focus on prediction tasks rather than cross-cultural aspects of wine appreciation.

Cross-cultural analysis of flavors. Flavor perception varies across cultures, as shown in studies analyzing beer pairing preferences in Latin America (Arellano-Covarrubias et al., 2019) and color-flavor associations in snack packaging across China, Colombia, and the UK (Velasco et al., 2014). Research on cheddar cheese flavor lexicons across different countries (Drake et al., 2005) further highlights cultural influences on taste perception. consumer conceptualisations of "red" and "white" diverge sharply between French, Portuguese and South-African drinkers (Fairbairn et al., 2024), while translation studies demonstrate how Western terroir terms acquire auspicious connotations in Chinese branding (Niu, 2023). These findings underscore the need for culturally adaptive approaches in wine description and translation.

Cultural adaptation. As culture and language are intertwined and inseparabled, there is a rising demand to equip machine translation systems with greater cultural awareness (Nitta, 1986; Ostler, 1999; Hershcovich et al., 2022). However, it is costly and time-consuming to collect culturally sensitive data and perform a human-centered evaluation (Liebling et al., 2022). Cultural adaptation in NLP focuses on adjusting text style while preserving its original meaning (Jin et al., 2022a). This is particularly important in cross-cultural tasks, where differences in rhetorical structures and communicative conventions influence translation. For culturally-aware translation, eliciting explanations has been shown to significantly enhance comprehension of culturally specific entities (Yao et al., 2024; Singh et al., 2024). A recent study showed that LLMs are adept at adapting cooking recipes across cultures, using a comparable and a parallel corpus (Cao et al., 2024). Our work focuses on the translation of non-parallel subjective reviews, an area that remains underexplored.

3 Data Collection & Analysis

We present CulturalWR, a dataset of paired professional wine reviews. Here we describe how we construct our dataset and perform several analysis to answer RQ1.

3.1 Data Collection

We collect the English wine reviews from one wine sales website Wine and a vertical search engine Wine-Searcher. Chinese wine reviews are primarily written by two prominent reviewers, AlexandreMa[2], ShenHao[3], who often publish bilingual wine reviews and China Wine Information Network[4].

3.1.1 Data preprocessing

Wine names are usually derived from regions or grape varieties, labeled by these features or a unique name. We observed that Chinese reviewers often skip mentioning the grape variety and region, opting for simpler names. We use string matching with character normalization (e.g., converting "Château Nénin" to "Chateau Nenin") to match reviews to specific wines. Exact matches are assumed to indicate the same wine; partial matches undergo manual verification.[5] After confirming the names and vintage years match, we finalize the reviews for each specific wine.

To focus on red wines and adapt reviews culturally, we applied three filtering rules: excluding Non-Vintage (N.V.) wines due to inconsistent tasting notes over time, separating white and rosé wines into distinct datasets for independent testing, and filtering out reviews under 30 words in English or 30 characters in Chinese for lack of detail.

3.2 Dataset Overview

We collect approximately 20k Chinese wine reviews covering nearly 20k wines, along with 150,000 English wine reviews spanning over 30k wines. These reviews encompass a variety of wine types, including red, white, rosé, and champagne. After filtering, we retain around 10k Chinese reviews focused on red wines. Through matching, the dataset is refined to include 4.5k wines, comprising about 4.5k Chinese reviews and 16k English

[2] https://www.alexandrema.com

[3] http://www.leparadisduvin.com

[4] www.winesinfo.com

[5] Particularly in the Bordeaux region, wines use the winery's name for the Grand Vin (first wine) and unique names for second and third wines, with "blanc" added for white wines.

reviews. 3,227 Chinese reviews have their matched English reviews written by the same author. Data statistics are shown in Table 1.

	Number	Mean #Tokens
CA Chinese Reviews	4776	67.57
CA English Reviews	3227	74.25
Transl. Chinese Reviews	60	60.2
WA English Reviews	16746	58.16

Table 1: Statistics of reviews. CA refers to Chinese wine critics and WA to Western ones. We count tokens with `jieba` text segmentation for Chinese and whitespace tokenization for English.

Attributes. Besides the basic reviews and their corresponding rating for each wine, we sourced wine data from different sources, we reorganize the basic attributes of all these wines, which includes the geographical location of the winery, the composition of the grape varietals, the vintage, the alcohol content and the price. To ensure privacy, we anonymized the obtained data, and no personally identifiable information is included in the attributes. See Appendix G.

4 Cross Cultural Wine Reviews Adaptation Task

We propose cross-cultural wine review adaptation, extending machine translation by requiring both accuracy and deliberate semantic divergence to address cultural differences.

Evaluating cultural adaptation is challenging, as it must balance meaning preservation with genuine cross-cultural differences. In wine review adaptation, we even need to adapt the flavor descriptors. As common in text generation tasks, we first adopt reference-based automatic evaluation metrics. Moreover, considering reference-based metrics are often unreliable for subjective tasks, we also conduct human evaluations.

4.1 Automatic Evaluation

We use four metrics to assess the similarity between the generated and reference reviews. We use two lexical-based metric: BLEU (Papineni et al., 2002), a precision metric based on token n-gram which emphasizes precision and commonly used in machine translation evaluation and METEOR (Banerjee and Lavie, 2005), which combines precision and recall while incorporating linguistic features such as stemming and synonymy to provide a more comprehensive evaluation; one contextual-embedding based metric: BERTScore (Zhang* et al., 2020), based on cosine similarity of contextualized token embeddings and capture deep semantic matching; one hybrid-based metric: Beer (Stanojević and Sima'an, 2014), based on multi-feature fusion regression indicator, which combines syntactic and semantic features to automatically evaluate translation quality through regression model learning.

4.2 Human Evaluation

While automatic metrics provide quantifiable results, they rely on fixed reference sets, which may lack cultural relevance. To address this, we introduce seven human evaluation criteria applied to the test set.

(1) **Grammar**: The generated reviews are grammatically correct and fluent; (2) **Faithfulness of Information**: The content accurately reflects the original input without introducing false or misleading information; (3) **Faithfulness of Style**: The output preserves the original tone, register, and formality without altering the intended voice; (4) **Overall quality**: The reviews are coherent, contextually appropriate, and align with the intended tone and style; (5) **Cultural Proximity**: The generated reviews use familiar terms and expressions that resonate with the target culture. For example black currant was replaced by Chuanbei loquat paste, cough syrup and hawthorn cake in Chinese localised aroma wheel (Jin et al., 2022b); (6) **Cultural Neutrality**: Maintains neutrality to avoid provoking negative perceptions or reactions from the target culture consumers. For example, 'earthy' can be a positive descriptor for wine in the West. However, when translated into Chinese, '土味' often implies a dirty or unrefined flavor. A more elegant term like '泥土气息' (aroma of soil) is often preferred; (7) **Cultural Genuineness**: Preserves the quality of the original descriptor without altering its meaning, ensuring authenticity. For example clove, violet and saffron have no suitable local descriptors with similar olfactory characteristics.

Our evaluation was done by five people fluent in both Chinese and English, including four master students from various majors [6] and a Stage 3 Master of Wine student. Before the evaluation process, we performed preliminary testing on 40 samples

[6] The annotators are from Computer Science, Food Science and Data Science.

and used Pearson correlation to calculate their understanding of different metrics which proved these metrics are easy for people to evaluate.

5 Cross Cultural Reviews Analysis

5.1 Lexical difference analysis

To answer **RQ1**, we first manually extracted detailed flavor descriptors from 250 Chinese and 250 Western wine reviews and map detailed flavor descriptors into three hierarchical layers: **Aroma Families** (broad categories such as fruits, flowers, and spices), **Aroma Subfamilies** (more specific groups like citrus fruits, red berries, and dried herbs), and **Exact Aromas** (precise descriptors such as lemon, raspberry, or thyme)—leveraging the Wine Aroma Wheel from Aromaster[7], with Chinese-localized descriptors proposed by (Jin et al., 2022b). We found that over 300 unique descriptors were identified in the corpus, although they totally offer about 100 flavor descriptors. To facilitate systematic cross-cultural comparison, we constructed a culturally aligned inventory of 92 standardized descriptors from the merged wheel, which includes flavors mentioned at least five times in the selected reviews. We attribute the variation in descriptor usage to individual author preferences rather than broadly shared sensory conventions. An example of how we applied this method can be found in Appendix J. Based on this mapped inventory, we then compute the inner- and outer-group Jaccard similarities[8] over these Aroma Hierarchies.

	Exact Aromas	Aroma Subfamilies	Aroma Families
Inner Sim	0.14	0.25	0.48
Outer Sim	0.08	0.18	0.40

Table 2: Inner- and Outer-Group Jaccard Similarity Across Aroma Hierarchies

From Table 2, we observe that at the Exact-Aroma level the average inner-group Jaccard similarity is 0.14, whereas the outer-group score falls to 0.08, revealing scant lexical overlap across cultures. At the Aroma-Subfamily level the outer similarity improves to 0.18, yet the gap to the inner score (0.25) persists, confirming noticeable divergence. Even at the most abstract tier—Aroma Families—the outer similarity climbs only to 0.40, still trailing the inner benchmark (0.48) and demonstrating that a sizeable cultural difference remains despite hierarchical mapping.

Overall, however, cross-cultural overlap remains modest even at the Aroma Families level, which we attribute to contrasting reviewing conventions: some critics mention only the dominant notes of complex wines, whereas others enumerate every perceivable aroma. And in both cultures, people still have their own tendencies toward specific flavor descriptors. The consistently lower outer-group scores highlight a clear need for cultural adaptation: without targeted lexical substitution, flavour descriptions that feel familiar to one readership may remain opaque—or even misleading—to another.

5.2 Semantic difference analysis

To further analyse semantic differences, we perform PCA on the review embeddings, which are extracted from the encoder layer of LaBSE(Feng et al., 2022). To mitigate distortion from dimensionality reduction and balance sample sizes, we uniformly sample 500 instances from each category. In addition, to minimise language-specific bias we focus on English reviews authored by Chinese reviewers. As shown in Figure 2, the two embedding clouds are almost separated in the reduced space, indicating substantial differences in the global structure and overall semantic distribution of the two review groups. When we instead contrast native Chinese-language reviews with English-language reviews, the separation widens even further, underscoring the strong confounding effect of surface language. The result can be seen in Appendix C. Synthesizing evidence from both the embedding-level and lexical level, Chinese and Western reviews remain essential differences even under strict controls, further reinforcing the need for cultural adaptation.

5.3 Cross-Cultural Comprehension Analysis

To answer **RQ2**, we randomly select 40 reviews from each culture and let three annotators label the initial untranslated reviews on Cultural Proximity and Cultural Neutrality [9].

As shown in Table 3, annotators tend to give higher scores to reviews from their own culture. Additionally, Western annotators consistently assign higher scores than Chinese annotators, sug-

[7] https://aromaster.com/

[8] "Outer" refers to comparisons between Chinese and Western reviews, while "Inner" refers to comparisons within Western reviews.

[9] Inner = ratings from annotators of the same culture; Outer = ratings from annotators of the different culture.

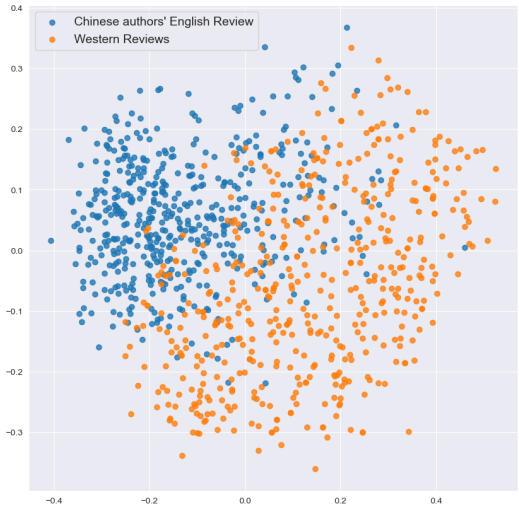


Figure 2: PCA-reduced embeddings: Blue dots show Chinese Authors' English reviews, and orange dots show Western reviews

Culture	Inner C-P	Outer C-P	Inner C-N	Outer C-N
Chinese	4.5	4.9	4.7	5.3
Western	6.2	5.6	6.7	5.3

Table 3: Cross-Cultural Ratings of Cultural Proximity and Neutrality in Chinese and Western Reviews

gesting that Western wine reviews may be more difficult for Chinese readers to understand, compared to how Western readers perceive Chinese wine reviews.

6 Experiment

To comprehensively assess the efficacy of LLM translations in understanding wine and flavors, we compare various prompting strategies on tuning-free LLMs, alongside evaluations of an open-source Machine Translation model.

6.1 Experimental Setup

Prompting LLMs. Based on the exceptional performance of multilingual LLMs in zero-shot translation. We explore their ability on translating wine reviews and flavor adaptation.

We compare the performance of five diverse, state-of-the-art LLMs on this task. We test Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023),Phi-3.5-mini-instruct (Abdin et al., 2024), and ChatGPT-4o (OpenAI et al., 2024) and two use more Chinese training data models: Qwen2.5-7B-Instruct (Yang et al., 2024), GLM4-9b (GLM et al., 2024). We used Cultural Prompt in Table 11.

Multilingual machine translation model: We use the state-of-the-art NLLB-200-3.3B model (Team et al., 2022) for accurate, context-aware multilingual translation in our experiments.

7 Results and Analysis

Our analysis includes five parts: 1) automatic evaluation between different models; 2) fine-grained human evaluation on a subset of Chinese-English translation bidirectional; 3) Correlation of automatic metrics with humans; 4) Different prompting strategy evaluation comparison; 5) Quantitative analysis for some specific concepts

7.1 Overall Automatic Evaluation

Models	BLEU	METEOR	B-Sc	BEER	#Tok
Chinese → English					
ChatGLM4	**18.7**	45.7	86.7	51.5	65.8
Phi	10.9	40.6	90.0	47.5	75.6
Qwen2.5	15.6	46.3	**91.2**	52.7	69.2
Mistral	5.2	36.4	87.9	34.1	61.8
Llama3.1	11.4	35.2	88.0	44.3	54.0
NLLB	12.0	36.8	88.7	48.1	64.8
ChatGPT	18.4	**48.5**	**91.2**	**53.4**	78.5
English → Chinese					
ChatGLM4	8.9	31.9	86.6	26.5	130.5
Phi	4.8	25.8	89.8	22.5	94.7
Qwen2.5	12.3	36.4	**90.6**	29.3	85.7
Mistral	2.0	20.1	89.3	17.0	56.1
Llama3.1	9.9	31.9	87.3	28.1	89.5
NLLB	3.8	18.3	87.6	21.4	96.8
ChatGPT	**16.4**	**41.1**	90.2	**34.0**	90.3

Table 4: Automated evaluation results on the test sets using reference-based metrics: BLEU, METEOR, B-Sc(BERTScore) and BEER. Higher scores indicate better performance on all metrics.

For Chinese-to-English translation, ChatGLM4 achieved the highest BLEU score, suggesting strong lexical matching, while ChatGPT excelled in BERTScore and BEER, indicating superior semantic alignment with human-written translations. For English-to-Chinese translation, ChatGPT led in BLEU and METEOR, and Qwen2.5 again outperformed in BERTScore, reinforcing its strength in preserving meaning. Notably, ChatGLM4 struggled with BLEU and METEOR in this direction but maintained competitive BERTScore values, suggesting it prioritizes semantic coherence over strict lexical overlap. Additionally, the NMT system NLLB still remains highly competitive. Overall, ChatGPT and Qwen2.5 consistently demonstrated high semantic quality across both directions, while other models exhibited strengths in specific metrics, reflecting differing optimization objectives. These

results highlight the inherent trade-offs between fluency, lexical fidelity, and semantic preservation across models. More importantly, they underscore that reference translations are not the sole "correct" adaptations, as translation quality is inherently subjective. This reinforces the need for a nuanced evaluation framework that accounts for cultural context, linguistic variation, and domain-specific preferences to better capture real-world translation quality.

7.2 Human Evaluation

Models	F-I	F-S	Gr	O-Q	C-P	C-G	C-N
Chinese → English							
ChatGLM4	5.6	4.8	5.3	5.0	6.8	**6.4**	6.2
Phi	4.8	5.4	6.0	4.8	6.9	6.3	**6.4**
Qwen2.5	**5.8**	**6.0**	5.8	**5.5**	6.8	6.3	**6.4**
Mistral	4.8	5.7	6.0	4.8	**7.0**	**6.4**	6.3
Llama3.1	5.3	5.6	**6.2**	5.3	**7.0**	6.2	6.3
Human	5.5	5.3	5.5	5.3	6.6	6.3	6.2
English → Chinese							
ChatGLM4	5.5	5.5	4.2	4.7	5.6	5.5	5.3
Phi	4.1	4.5	3.6	3.7	5.6	4.4	5.3
Qwen2.5	**5.7**	**5.9**	**5.5**	**5.6**	**6.0**	5.5	**5.8**
Mistral	3.8	4.4	2.8	3.1	4.8	4.6	4.7
Llama3.1	4.5	5.0	4.4	4.6	5.9	4.8	5.7
Human	5.1	5.8	5.2	5.0	5.8	**5.7**	**5.8**

Table 5: Human evaluation results on the selected test sets: average for each method and metric, ranging from 1 to 7. F-I, F-S, Gr, O-Q, C-P, C-G and C-N representing Faithful of Information, Faithful of Style, Grammar, Overall Quality, Cultural Proximity, Cultural Genuineness and Cultural Neutrality respectively

Table 5 presents the results of human evaluation across multiple dimensions. Notably, NLLB was excluded from human evaluation due to its lack of cultural adaptation capabilities.

For English-to-Chinese translation, Qwen2.5 leads across most metrics, even surpassing human translations in faithfulness, grammar, and overall quality. Human translations score highest in Cultural Genuineness, reflecting their ability to preserve nuanced expressions. Llama3.1 also remains competitive, balancing fluency and cultural adaptation.

For Chinese-to-English translation, Qwen2.5 again ranks highest in faithfulness and overall quality, while Mistral outperforms even human translations in Cultural Proximity and Genuineness, suggesting a stronger emphasis on natural, idiomatic English.

These results highlight trade-offs between literal accuracy and cultural adaptation. While human translations excel in cultural authenticity, LLMs like Qwen2.5 demonstrate strong faithfulness, and Mistral prioritizes fluency in the target language. This underscores the importance of context-aware evaluation that considers both linguistic accuracy and cultural nuances.

For the feedback from the Stage 3 Master of Wine student, we still find that Model ChatGLM has correct vocabulary but has poor grammar, also very literal. Phi has common wrong translations, but has fixed wine tasting vocabulary. Llama is the most metaphorical and tend to reword stuff. Mistral just skips proper nouns but eloquent. ChatGLM, ChatGPT and Human translation are best. Human translation and ChatGPT are maybe too literal (correct but boring), ChatGLM are super eloquent.

Methods	F-I	F-S	Gr	O-Q	C-P	C-G	C-N
Chinese → English							
Human-Eval	5.36	5.58	5.91	5.19	6.86	6.31	6.36
GPT4o-Eval	5.79	5.24	6.47	5.77	5.23	5.4	6.44
English → Chinese							
Human-Eval	4.57	5.02	4.44	4.13	5.25	5.12	5.0
GPT4o-Eval	5.76	5.36	6.2	5.71	5.39	5.58	6.43

Table 6: GPT-4o v.s. Human in Human evaluation metrics (Average over All Human Test Cases)

GPT-4o and Human Evaluation Diverge. Table 6 shows the results evaluated by both GPT-4o and human annotators. Specifically, we use all the human evaluation test cases for GPT evaluation. The results show that GPT and Human evaluation scores are not strongly correlated, and GPT has a higher tolerance than humans, especially for the English → Chinese direction. This discrepancy suggests that GPT-4o may have a different interpretation of translation quality compared to human evaluators. This discrepancy is particularly evident in culturally relevant metrics. The prompts we used are shown in Appendix D.

7.3 Correlation of automatic metrics with humans

To evaluate the reliability of automatic metrics for wine review adaptations, we analyze their correlation with human evaluations across seven metrics using Kendall correlation, the WMT22 meta-evaluation standard (Freitag et al., 2022).

As illustrated in Table 7, the correlation between human evaluation results and automatic metrics varies across different translation directions and evaluation criteria. For Chinese → English, F-I (Faithful of Information) and O-Q (Overall Qual-

	BLEU	METEOR	B-Sc	BEER
Chinese → English				
F-I	0.2536*	0.1892	0.3000*	0.2442*
F-S	0.0053	0.0075	0.0204	0.0113
Gr	-0.0296	-0.0096	0.0033	-0.0478
O-Q	0.2191*	0.1847*	0.2061*	0.2015*
C-P	-0.070	-0.0177	-0.0540	-0.0961
C-G	0.1131	0.0625	0.0621	0.1131
C-N	-0.1166	-0.0841	-0.0166	-0.0876
English → Chinese				
F-I	0.4079*	0.2788*	0.4080*	0.2946*
F-S	0.3723*	0.3560	0.3769*	0.3904*
Gr	0.2819*	0.2788*	0.3530*	0.2946*
O-Q	0.3526*	0.3123*	0.3742*	0.3557*
C-P	0.1408	0.1524	0.2609*	0.1553
C-G	0.3134*	0.3170*	0.3942*	0.3170*
C-N	0.1334	0.1357	0.2389*	0.1516

Table 7: 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

ity) exhibit the strongest correlations, particularly with BLEU, B-Sc, and BEER, suggesting that these metrics align well with human judgments in assessing fluency and overall translation quality. On the other hand, for English-Chinese, the correlations are generally stronger across all metrics. Notably, F-I, F-S (Faithful of Style), and O-Q display significant correlations with multiple metrics, particularly B-Sc and BEER, indicating that these metrics are relatively more reliable for evaluating fluency and intelligibility in English-to-Chinese translations.

C-G (Cultural Genuineness) also achieves strong correlations, indicating that automatic metrics can reliably assess translation accuracy. However, this does not necessarily reflect the cultural adaptability of the translation. However, C-N(Cultural Neutrality) and C-P(Cultural Proximity) remain weakly correlated, revealing that automatic metrics still fall short in capturing deeper cultural nuances, emphasizing the need for human evaluation. Notably, correlations for English→Chinese generally exhibit greater strength than Chinese→English. This discrepancy is likely due to most wine reviews being written from a predominantly Western reviewer's perspective.

7.4 Prompting Strategy Evaluation

With LLM-based machine translation advancing, integrating free-form external knowledge offers new opportunities to enhance translation quality.

Human Evaluation							
Methods	**F-I**	**F-S**	**Gr**	**O-Q**	**C-P**	**C-G**	**C-N**
Direct Translation	6.38	5.94	5.56	5.69	4.38	6.31	5.81
Cultural Prompt	6.25	6.12	5.88	5.94	4.5	6.19	5.94
Detailed Cultural Prompt	5.75	5.94	5.75	5.56	4.88	5.69	5.69
Self-Explanation	4.94	5.75	5.88	5.5	5.75	4.56	6.62
Automatic Evaluation							
Methods	**BLEU**	**METEOR**	**B-Sc**	**BEER**			
Direct Translation	16.5	41.6	91.4	28.3			
Cultural Prompt	15.9	41.0	91.3	28.3			
Detailed Cultural Prompt	14.4	38.6	91.0	27.6			
Self-Explanation	13.7	37.0	90.9	27.6			

Table 8: Evaluation of different strategies by GPT4o on English-Chinese translations.

We compare different prompting strategies on ChatGPT for English-to-Chinese translation, including Direct Translation, Cultural Prompt, Detailed Cultural Prompt, and Self-Explanation. Table 11 lists the specific prompts.

As is shown in Table 8, Direct Translation is the best-performing method in Faithful of Information and automatic metrics. Cultural Prompting improves cultural accuracy slightly but does not significantly enhance overall translation quality. Self-Explanation has the worst faithful scores but is rated the best in Cultural Neutrality, making it a trade-off strategy for culturally rich contexts. This trade-off suggests that translation strategies need to balance faith, accuracy, and cultural adaptability, depending on the intended use case.

8 Conclusion

In this work, we studied cross-cultural adaptation of wine reviews, introducing CulturalWR, a dataset of paired Chinese and English reviews, and evaluating LLM-based adaptation methods. Our results show that LLMs can consider cultural nuances but face challenges in maintaining detail and consistency in flavor descriptions. We also assessed adapted flavor similarities to gauge LLMs' understanding of wine descriptors. Beyond wine reviews, our findings have broader implications for cross-cultural communication in the wine industry, aiding wineries and retailers in tailoring descriptions to international audiences. This could enhance global wine marketing while preserving cultural authenticity. Moreover, our work highlights AI's potential in gastronomy, paving the way for research into AI-assisted flavor profiling and food pairing. Future work includes refining adaptation models for coherence, integrating multimodal data, and using user feedback to enhance AI-generated adaptations, bridging cultural gaps in wine appreciation.

9 Limitation

Cultural adaptation in wine reviews has great potential to aid consumer decisions and promote wines globally. However, several challenges remain. A key limitation is the dataset size and diversity. Our study includes fewer than 5,000 Chinese reviews, primarily from three professional critics, raising concerns about representativeness. Additionally, while our dataset contains reviews in multiple languages (German, Portuguese, French, Dutch, Italian, and Spanish), we focused solely on Chinese and English due to resource constraints. Expanding multilingual analysis could offer further insights. Another constraint lies in evaluation prompts. Our analysis relies on four prompts, which may not fully capture how models handle cultural adaptation. Broader prompt variations and real-world user inputs could improve evaluation robustness. Moreover, our quality assessment is based on the Wine Aroma Wheel and prior sensory adaptation research (Jin et al., 2022b), but it lacks independent verification of adaptation accuracy. Future work could incorporate human or expert evaluations for a more comprehensive assessment. Finally, wine reviews are inherently subjective, influenced by personal preferences and sensory perceptions. While we strive to minimize bias, eliminating subjectivity entirely remains challenging. Addressing this may require larger, more diverse datasets and structured sensory evaluation frameworks in future research.

Ethical Considerations

This study relies on wine ratings and tasting notes sourced from professional websites, all of the datasets are all publicly accessible and do not involve any user privacy. Our experiments strictly comply with the terms of use of these platforms

To support transparency and reproducibility, we have released all source code and experimental configurations. In addition, we have publicly shared the dataset that does not contain proprietary content and can be accessed through open-access sources, subject to applicable licensing terms. An anonymous link to the code and data subset is provided for review purposes and will be replaced with a permanent repository upon publication. We have licensed our data under CC0 1.0 DEED and will ask the users to sign an agreement such that the dataset will only be available for academic research purposes to further protect the users.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Araceli Arellano-Covarrubias, Carlos Gómez-Corona, Paula Varela, and Héctor B. Escalona-Buendía. 2019. Connecting flavors in social media: A cross cultural study with beer pairing. *Food Research International*, 115:303–310.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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.

Bernard Chen, Christopher Rhodes, Alexander Yu, and

Valentin Velchev. 2016. The computational wine wheel 2.0 and the trimax triclustering in wineinformatics. In *Advances in Data Mining. Applications and Theoretical Aspects*, pages 223–238, Cham. Springer International Publishing.

Decanter China. 2023. China refreshes its record of medal wins at Decanter World Wine Awards 2023. Accessed: 2025-05-13.

Jyh-Shen Chiou, Cheng-Chieh Hsiao, and Fang-Yi Su. 2014. Whose online reviews have the most influences on consumers in cultural offerings? professional vs consumer commentators. *Internet Research*, 24(3):353–368.

Lukas Danner, Trent E Johnson, Renata Ristic, Herbert L Meiselman, and Susan EP Bastian. 2017. "i like the sound of that!" wine descriptions influence consumers' expectations, liking, emotions and willingness to pay for australian white wines. *Food Research International*, 99:263–274.

M.A. Drake, M.D. Yates, P.D. Gerard, C.M. Delahunty, E.M. Sheehan, R.P. Turnbull, and T.M. Dodds. 2005. Comparison of differences between lexicons for descriptive analysis of cheddar cheese flavour in ireland, new zealand, and the united states of america. *International Dairy Journal*, 15(5):473–483.

Samantha Fairbairn, Jeanne Brand, Antonio Silva Ferreira, Dominique Valentin, and Florian Bauer. 2024. Cultural differences in wine conceptualization among consumers in france, portugal and south africa. *Scientific Reports*, 14:15977.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

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.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Thomas Hörberg, Murathan Kurfalı, and Jonas K. Olofsson. 2025. Chemosensory vocabulary in wine, perfume and food product reviews: Insights from language modeling. *Food Quality and Preference*, 124:105357.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022a. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.

Gang Jin, Xi Lv, Linsheng Wei, Laichao Xu, Junxiang Zhang, Yanping Chen, and MA Wen. 2022b. Chinese localisation of wine aroma descriptors: an update of le nez du vin terminology by survey, descriptive analysis and similarity test. *OENO One*, 56(1):241–251.

Els Lefever, Iris Hendrickx, Ilja Croijmans, Antal van den Bosch, and Asifa Majid. 2018. Discovering the language of wine reviews: A text mining account. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Yoshihiko Nitta. 1986. Problems of machine translation systems: Effect of cultural differences on sentence structure. *Future Generation Computer Systems*, 2(2):101–115.

Ning Niu. 2023. Terroir, wine varieties, brands and their translations into chinese. *International Journal of Chinese and English Translation & Interpreting*, 2(2).

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi,

Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Nicholas Ostler. 1999. "the limits of my language mean the limits of my world": is machine translation a cultural threat to anyone? In *Proceedings of the 8th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, University College, Chester.

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.

Heber Rodrigues and Wendy V Parr. 2019. Contribution of cross-cultural studies to understanding wine appreciation: A review. *Food research international*, 115:251–258.

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. Translating across cultures: Llms for intralingual cultural adaptation. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, page 400–418. Association for Computational Linguistics.

Charles Spence. 2020. Wine psychology: basic & applied. *Cognitive research: principles and implications*, 5(1):22.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA. Association for Computational Linguistics.

Alejandro Suárez, Heber Rodrigues, Samantha Williams, Purificación Fernández-Zurbano, Nicolas Depetris-Chauvin, Gregory Dunn, and María-Pilar Sáenz-Navajas. 2023. Effect of culture and familiarity on wine perception: A study with british and spanish wine experts. *OENO One*, 57(2):61–69.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Carlos Velasco, Xiaoang Wan, Alejandro Salgado-Montejo, Andy Woods, Gonzalo Andrés Oñate, Bingbing Mu, and Charles Spence. 2014. The context of colour–flavour associations in crisps packaging: A cross-cultural study comparing chinese, colombian, and british consumers. *Food Quality and Preference*, 38:49–57.

Vincent Xian Wang, Xi Chen, Songnan Quan, and Chu-Ren Huang. 2021. A parallel corpus-driven approach to bilingual oenology term banks: How culture differences influence wine tasting terms. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

Wine Intelligence. 2024. China's wine exports to Japan soar by 255% in 2023. Accessed: 2025-05-13.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. *Preprint*, arXiv:2305.14328.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A French and Spanish Character Conversion Table

This section shows the Character Conversion Table we have used to convert some characters, shown in Table 9.

B Whole dataset of Other Languages

It's the all the data we have collected, besides Chinese and English, we also collected some reviews written in German, Portuguese, French, Dutch, Italian and Spanish.

Original	Converted	Original	Converted
à	a	â	a
ä	a	é	e
è	e	ê	e
ë	e	î	i
ï	i	ô	o
ö	o	ù	u
û	u	ü	u
ç	c	ÿ	y
æ	ae	œ	oe
À	A	Â	A
Ä	A	É	E
È	E	Ê	E
Ë	E	Î	I
Ï	I	Ô	O
Ö	O	Ù	U
Û	U	Ü	U
Ç	C	Ÿ	Y
á	a	í	i
ó	o	ú	u
ñ	n	Á	A
Í	I	Ó	O
Ú	U	Ñ	N

Table 9: French and Spanish Character Mapping to English

	Numbers	Mean Tokens
CA Chinese Reviews	4776	67.57
CA English Reviews	3227	74.25
WA English Reviews	16746	58.16
WA German Reviews	3341	50.73
WA Portuguese Reviews	161	82.15
WA French Reviews	480	135.63
WA Italian Reviews	40	44.35
WA Spanish Reviews	10	114.8

Table 10: Statistics of reviews. We count tokens with jieba text segmentation for Chinese and whitespace tokenization for other languages.

C Embedding Visualization

Figure 3: PCA-reduced embeddings: Blue dots show Chinese reviews, and orange dots show Western reviews

D Prompt Used for Evaluation

Table 11 shows different prompt strategies we used for Prompting Strategy Evaluation.

Here shows the prompt we used for GPT evaluation:

```
As a Western consumer, evaluate the
quality of wine review translation
from these dimensions, use seven-tier
scoring ranging from 1-7 : 7 means
Excellent, 6 means Very Good, 5
means Good, 4 means Fair, 3 means
Poor, 2 means Very Poor, 1 means
Nonsense: Grammar,Faithfulness of
Information,Faithfulness of Style,
Overall quality,Cultural proximity
- The generated reviews use
familiar terms and expressions that
resonate with the target culture,
Cultural neutrality - Maintains
neutrality to avoid provoking
negative perceptions or reactions
from the target culture consumers,
Cultural Genuineness - Preserves
the quality of the original
descriptor without altering its
meaning, ensuring authenticity.
Original: {original}
Translation: {translation}
```

E Grading scale rule for human evaluation

Here shows the rule for the Seven-tier rating system:

1. Excellent (rating: 7 points)
   The translation is also highly matched in context and culture.

2. Very good (rating: 6 points)
   There may be slight (1-2) inaccuracy of vocabulary or inadequate reflection of some details, but it does not affect the overall understanding.

Strategy	Prompt
Direct Translation	Translate the following English wine reviews to Chinese: [Wine review]
Cultural Prompt	Translate the wine reviews in Chinese, adapted to an Chinese-speaking consumer: [Wine review]
Detailed Cultural Prompt	Translate the provided English wine review into Chinese, so that it fits within Chinese wine culture and to avoid using any terms that might have negative connotations for Chinese consumers: [Wine review]
Self-Explanation	User: Find flavor and aroma descriptions that are unfamiliar and uncomfortable for Chinese consumers: [Wine review] LLM: [Sentences] User: Translate the wine review in Chinese, and for the unfamiliar and uncomfortable flavor and aroma, replace it with a more familiar and comfortable description for Chinese consumers.

Table 11: Prompting strategy examples used for English → Chinese translation

Strategy	Prompt
Direct Translation	Translate the following Chinese wine reviews to English: [Wine review]
Cultural Prompt	Translate the wine reviews in English, adapted to an Western-speaking consumer: [Wine review]
Detailed Cultural Prompt	Translate the provided Chinese wine review into English, so that it fits within Western wine culture and to avoid using any terms that might have negative connotations for Western consumers: [Wine review]
Self-Explanation	User: Find flavor and aroma descriptions that are unfamiliar and uncomfortable for Western consumers: [Wine review] LLM: [Sentences] User: Translate the wine review in English, and for the unfamiliar and uncomfortable flavor and aroma, replace it with a more familiar and comfortable description for Western consumers.

Table 12: Prompting strategy examples used for Chinese → English translation

3. Good (rating: 5 points)
   There are individual mistranslations, but they will not seriously change the overall meaning.

4. Fair (score: 4 points)
   Mistranslations are obvious, which may cause some semantics to deviate from the original text.

5. Poor (score: 3 points)
   The overall translation quality is low and may mislead readers.

6. Very poor (score: 2 points)
   It is basically impossible to rely on the translation to understand the original text.

7. Nonsense (score: 1 point)
   The translation is completely incoherent semantically and unable to convey the information which the original review tries to convey.

F Human evaluation platform

Figure 4 shows a screenshot from our human evaluation platform, demonstrating the English to Chinese direction. Human need to evaluation the quality of Chinese translation

G Attributes of Whole dataset

Figure 5 shows attributes counted in CulturalWR dataset.

H Experiment Settings

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

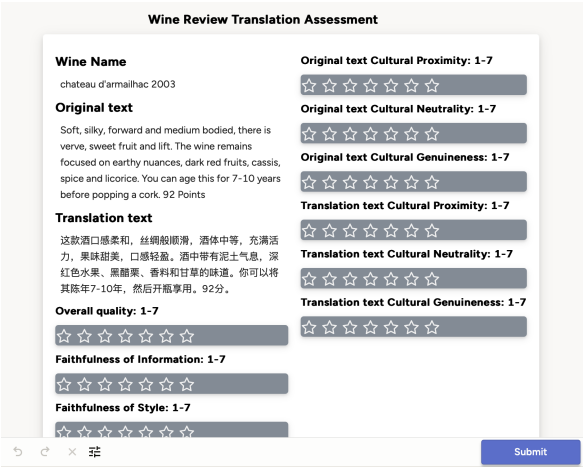
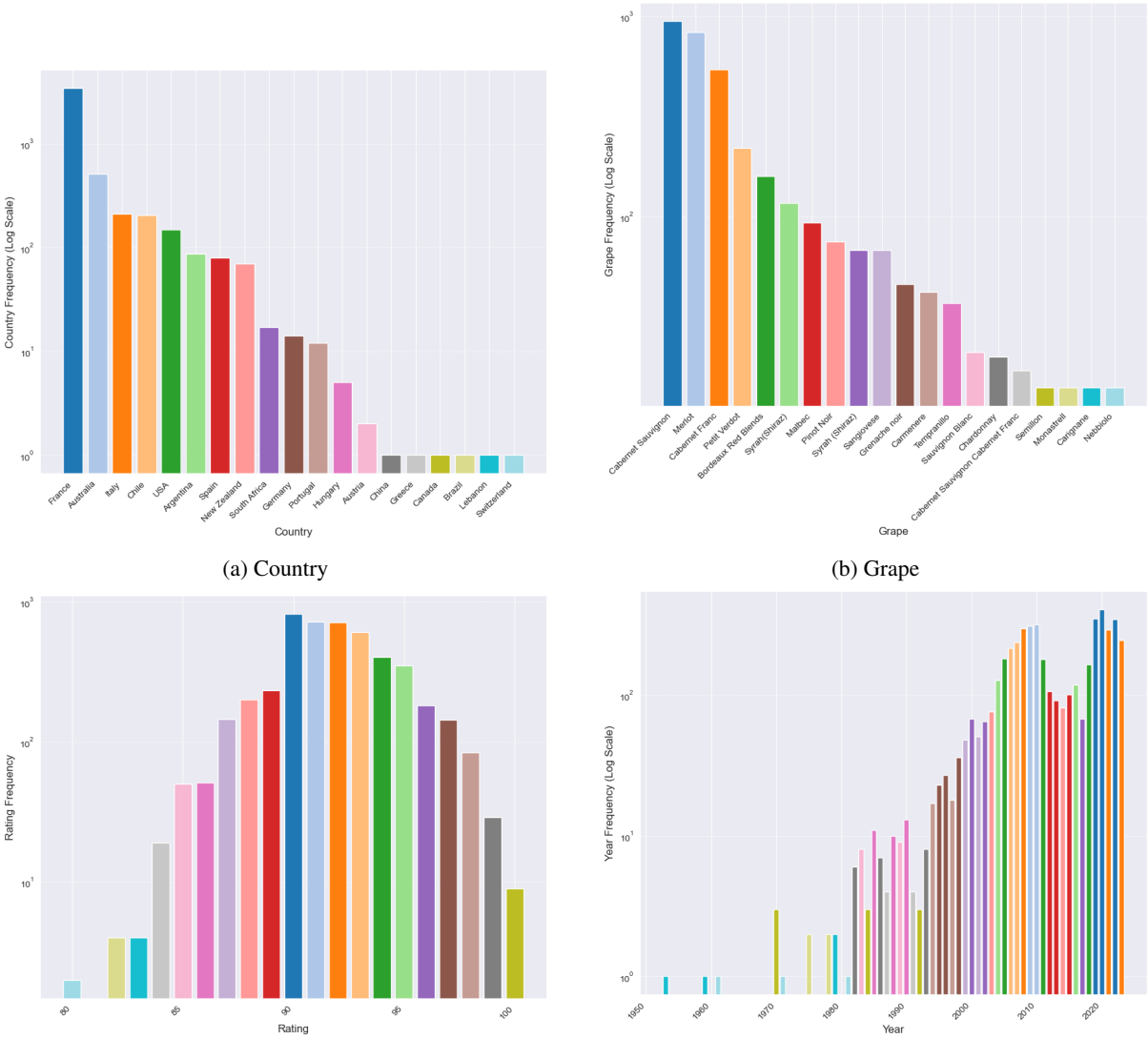


Figure 4: Screenshot from our human evaluation platform

1. **NLLB** We use NLLB-200-3.3B[10] for our experiments. The beam is set as 4, and the length penalty is set as 1.0.

2. **Pretrained LLMs** We used LLMs including Llama-3.1-8B-Instruct [11], Mistral-7B-Instruct-v0.3 [12], Phi-3.5-mini-instruct[13], Qwen2.5-7B-Instruct[14], GLM4-9b[15]. The

[10] https://huggingface.co/facebook/nllb-200-3.3B
[11] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[12] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[13] https://huggingface.co/microsoft/Phi-3.5-mini-instruct
[14] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[15] https://huggingface.co/THUDM/glm-4-9b



(a) Country

(b) Grape

(c) Rating

(d) Year

Figure 5: Wine attributes.

sampling is set as True, leading to a multinomial sampling searching method. All settings are the same across different models.

3. **ChatGPT** We used the latest version, GPT-4o-2024-11-20, through the ChatCompletion API provided by OpenAI [16]. For the generation, we set the parameters as default, for which the temperature is 1, top_p is 1, and frequency_penalty as 0.

I Examples

In Table 13, we present a comparative analysis of multiple machine translation models applied to an English-to-Chinese translation example which is shown in Figure 1. Each row in the table represents the output of a different model, with the original English input provided for reference. To facilitate an in-depth evaluation, we systematically annotate various translation characteristics, including direct translations, adapted phrasings, and potential errors.

To facilitate evaluation, we use color coding to distinguish different translation characteristics: red indicates incorrect translations, teal represents accurate literal translations, blue highlights adapted translations that maintain meaning while improving fluency, and brown marks uncertain adaptations.

This comparison reveals variations in how different models interpret and translate key terms, particularly in handling domain-specific vocabulary such as "Merlot" and "Mediterranean scrub." Some

[16] https://platform.openai.com/docs/guides/text-generation

Category	Content
Input (English)	Aromas of iris, raspberry, camphor and Mediterranean scrub mingle with oak-driven spice on this 100% Merlot.
ChatGLM4	这款酒散发着紫罗兰、树莓、樟脑和地中海灌木的香气，100%的梅洛。
Phi	这100%的梅鹿酒充满了茉莉花、覆盆子、茴香和地中海疏林的香气，与橡木风味相结合的精髓。
Qwen2.5	这款100%梅洛红酒散发出鸢尾花、覆盆子、薄荷和地中海灌木丛的香气，与橡木带来的香料味交织在一起。
Mistral	这是一款百分之一的梅洛酒，浓郁而舒服，携带了芙蓉花、莓果、樟木和植物的香气，以及橡木带来的香氛。
Llama	这款100%梅洛特红酒的香气中融合了百合花、草莓、樟脑和地中海灌木丛的气息，伴随着橡木驱动的香料。
NLLB	红虹，树，和地中海洗刷的香气与木驱动的香料混合在这个100%的梅罗特酒上。
ChatGPT	这款100%梅洛红酒散发出鸢尾花、覆盆子、樟脑和地中海灌木的香气，并与橡木带来的香料味交织在一起。

Table 13: Comparison of Translations from Different Models. Red text shows the wrong translation, teal text shows the correct literal translation, blue text shows the adapted translation and brown text shows unsure adapted translation.

models exhibit direct translation errors, while others apply adaptive strategies to enhance readability. By analyzing these differences, we can better understand the strengths and limitations of current machine translation systems.

J Jaccard Similarity Analysis Example

'Fully developed, the nose, with its spice box, tobacco leaf, forest floor, and cigar wrapper calls your name. Medium-bodied, fresh, vibrant, round, and packed with dark currants and red plums, this is ready to go. If you have a bottle, there is no reason to hold this any longer. It will not improve.'

For the above review, we manually extract the Exact Aromas [spices, tobacco, plum, black currant] which can easily be done. And then we use The Wine Aroma Wheel By Aromaster mapping them to [spices, toasted, stone fruit, red berries](Aroma Subfamilies) and [maturation for oak barry, Fruity red Wine]. Here spices and toasted all mapped in [maturation for 'oak barry and stone fruit', red berries all mapped in 'Fruity red Wine'. And based on these three sets we compute Jaccard similarity.

K Quantitative analysis

Cross-lingual translation of culturally specific concepts is challenging, as it requires balancing linguistic accuracy with cultural adaptation. Many models tend to rely on literal translation, which may not always convey the intended meaning naturally. To examine this, we evaluate each model's literal translation rate for culturally embedded flavor descriptors from the CulturalWR test set. For instance, in English-to-Chinese translation, 'thyme' is considered an English-specific concept. We count occurrences of related terms such as 'gooseberry', 'thyme', and 'rosemary' in English wine reviews (c_{source}) and record how often they are directly translated in the corresponding Chinese reviews (c_{target}) from model predictions. The literal translation rate is then calculated as $\frac{c_{target}}{c_{source}}$. To further assess translation quality, we conduct a bidirectional test on the five most common wine-related terms. This ensures that models not only preserve culturally specific terms when translating in one direction but also effectively map key wine descriptors between languages. Comparing accuracy across models provides insights into their ability to maintain semantic fidelity, crucial for expert-level wine translations.

As shown in Figure 6, we analyze six culturally

relevant concepts, three common in Chinese culture and three in Western culture. Results show significant differences in literal translation rates across models ChatGLM and Qwen, with a stronger emphasis on Chinese-language data, exhibit higher literal translation rates, prioritizing structural fidelity over adaptation. For Western cultural concepts, Llama and Mistral show some degree of accurate literal translation, though performance varies. Notably, no model directly translates 'waxberry', likely due to its regional specificity and the absence of a widely recognized equivalent in Western languages. NLLB struggles with all six concepts, highlighting potential NMT limitations. Furthermore, while Llama and Mistral do not achieve the highest literal translation rates, they demonstrate a tendency to adapt culturally specific terms rather than translate them directly. For example, they often map 'raspberry' to other red berries such as 'blueberry' and 'blackberry' and replace 'thyme' with similar spices like 'bay leaves' and 'cinnamon'. These adaptations align with the categorization in the Wine Aroma Wheel, as both the substituted berries and spices belong to the same Aroma Subfamilies. These findings are consistent with Table 5, where ChatGLM and Qwen score higher in Faithfulness of Information, while Llama and Mistral perform better in Cultural Proximity.

We further assess how well models handle wine terminology, which often has industry-specific meanings distinct from general usage. As shown in Figure 7, ChatGLM and Qwen perform well in translating these terms, while Phi also ranks highly, even leading for 'full'. However, challenges arise with terms such as 'nose', which refers to a wine's aroma in professional contexts. The phrase "on the nose" describes a wine's bouquet, yet most models translate 'nose' directly, failing to capture its specialized meaning. This highlights the challenge of translating wine-specific terms beyond their literal meanings and underscores the importance of integrating domain knowledge into machine translation models.

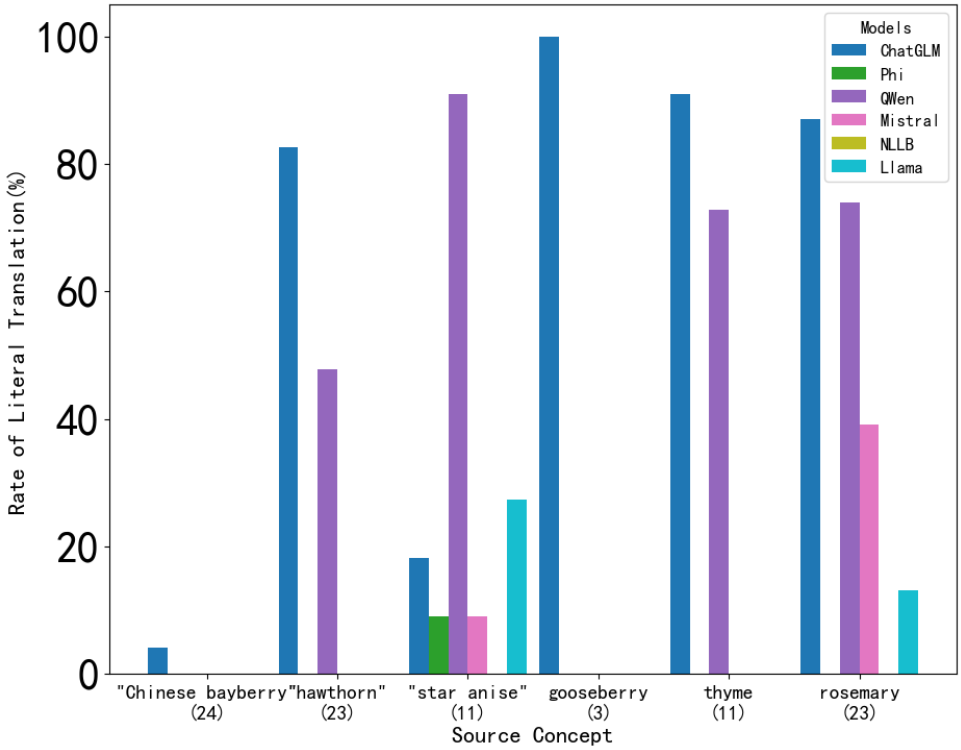


Figure 6: Analysis of the translation of specific concepts by the different models on the test data.

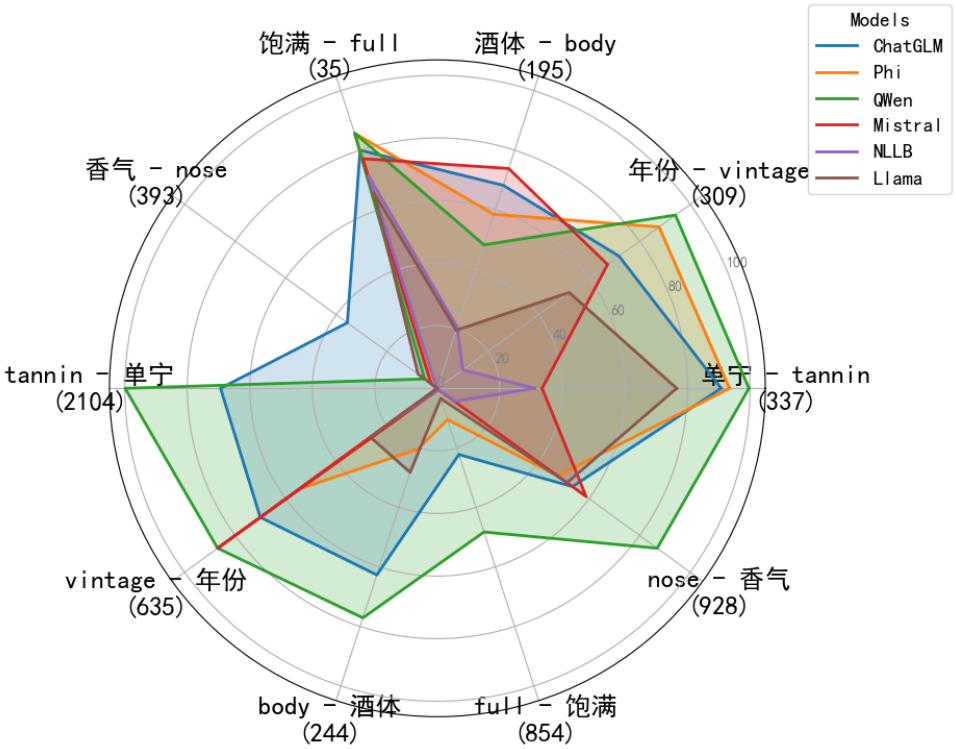


Figure 7: Analysis of the translation of specific terminology by the different models on the test data.