Evaluating Large Language Models on Chinese Idioms

Anonymous ACL submission

Abstract

Idioms are commonly used in everyday language, especially in Chinese, yet they pose a challenge to machine translation as their figurative meaning often differs from their literal parts. However, little is known about Chinese idiom translation, which poses unique challenges due to its semantic and pragmatic features and wide usage of proverbs. In this work, we introduce IDIOMEVAL, a comprehensive error taxonomy for Chinese idiom translation. We annotate 500 translation pairs from modern systems, including GPT-4 and Google Translate, across four domains: web, news, Wikipedia, and social media. Despite recent advances in neural machine translation powered by large language models, we find that these systems struggle significantly with idiom translation, frequently producing incorrect, literal, partial, or even missing translations. The best-performing system, GPT-4, introduces errors in more than 25% of idiom translations. Additionally, we evaluate how well existing automatic evaluation metrics measure idiom translation quality. We find these metrics have a low correlation with human ratings (with a Pearson correlation less than 0.53) and fail to reliably distinguish between good and bad idiom translations, achieving only 80% accuracy.

1 Introduction

Idioms are phrases whose meanings typically differ from the literal meanings of their individual components. They are commonly used in everyday language, making their understanding crucial for tasks such as neural machine translation (Baziotis et al., 2023) and writing assistance (Tan and Jiang, 2021; Dankers et al., 2022a). In particular, idioms have posed a challenge to translation as their meaning is usually non-compositional and cannot be inferred from their literal parts. For instance, *my two cents* means *one's viewpoint*, rather than actual money. Increasing attention has been paid to idiom translation in recent years. Dankers et al. (2022b) has found that transformer encoders tend to process figurative idioms as single units more strongly, while Baziotis et al. (2023); Liu et al. (2023) have shown different training methods can be helpful for idiom translation.

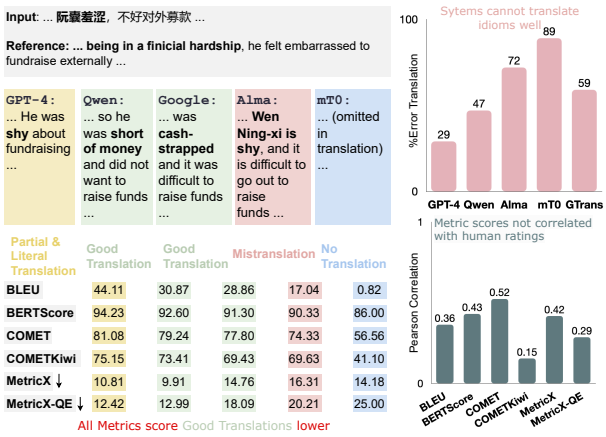

Figure 1: Top left: example of Chinese input and translations from five systems. Top right: percentage of good translations in each system. Bottom left: evaluation metric scores of system translations. Bottom right: Pearson's r between human rating and metric scores.

However, much of this research has focused on idioms in English, Japanese, and European languages, with little attention given to Chinese idioms. From a linguistics perspective, the former falls into the Indo-European language family, while Chinese belongs to the Sino-Tibetan language family (Katzner and Miller, 2002). Besides, idioms across languages differ in their definitions and classifications. For example, while English idioms often emphasize syntactic structures, Chinese idioms place more importance on semantic and pragmatic aspects (Wang, 2021). Meanwhile, Chinese idioms generally contain proverbs, allegorical sayings, and aphorisms which mostly originate from ancient literature (Wang, 2021, 2022). One example is the idiom 明日黄花, which first appeared in the Song dynasty and is literally translated into "tomorrow's

Error Type	Description	Example (zh → en)	
No Error	Translation is of high quality.	...科学家们欢呼雀跃，却发现...	...Scientists cheered and rejoiced, but they discovered that...
		...油霸们再捅刀美元，项庄舞剑意在沛公...	...oil tycoons stab the US dollar again, Xiang Zhuang's sword dance is aimed at Pei Gong...
Mistranslation	Translation is incorrect and disrupts idiom understanding.	...问得太多只会自寻烦恼...	Asking too many questions will only bring them problems. (*bring yourself problems*)
Unnatural	Translation is understandable yet suboptimal, and it can be improved.	...善战者无赫赫之功...	A good warrior does not achieve a glaring victory... (*a remarkable victory*)
Literal	Translation is literal and coherent with the surrounding context.	...这一生身无长处...	...I have no merits in this life... (*live in poverty*)
Addition	Translation contains irrelevant content beyond the accurate meaning.	...不仅于剧学素有深造，无所不通...	...not only well-versed in drama studies, knowledgeable and passionate in all aspects... (*knowledgeable in all aspects*)
Partial	Given idiom is translated partially.	...此处不留人，自有留人处，对于这类无良民宿...	...there is no place to stay, and against these bad inns... (*if you don't stay here, there is a place to stay elsewhere*)
Repetition	The translation contains repeated correct content or synonyms.	...最佳的父子关系应该如兄如弟...	...the best relationship between a father and son should be like that of a brother and a brother... (*like that of brothers*)
No Translation	Translation lacks the idiom's meaning in the output.	...正从布隆泉飞向普利托利亚，除了不速之客的蛇外，机上还载了4名乘客。	...was flying from Bloemfontein to Pretoria and had four four other passengers on board, including the snake. (the *unwelcome* snake)

Table 1: The IDIOMEVAL error taxonomy with example translations with each error type taken from our collected dataset in Section 3.2. The corrected span is included in *italics*.

yellow flower", but its figurative meaning is "things that have become outdated". Due to cultural and linguistic differences (Wang, 2022), Chinese idioms present unique challenges for neural translation systems.

In this work, we aim to shed light on how effectively modern translation systems handle Chinese idioms and whether current automatic metrics accurately capture the quality of idiom translations. We first introduce IDIOMEVAL, an error taxonomy capturing 7 different failure modes for Chinese idiom translation such as Mistranslation and Partial Translation (see Table 1. To explore the impact of domain context, we collect the Chinese texts that contain idioms from four distinct domains: Web, News, Wikipedia, and Social Media. Using the IDIOMEVAL taxonomy, we collect annotations for 500 translation pairs from 5 modern translation systems including GPT-4 (Achiam et al., 2023), Google Translate and several other large language models (LLMs). With these annotations, we conduct an error analysis on system translations and investigate how well current evaluation metrics assess the quality of Chinese idiom translations. Our main findings are as follows:

- The selected systems struggle to produce high-quality idiom translations. Their performance varies by domain, where they perform best on `News` but worst on `Social Media`.
- Translation errors differ across systems, but Mistranslation and Partial Translation are commonly seen, making up 54% of errors across all systems.
- Current machine translation evaluation metrics do not correlate well with human annotations.
- Existing metrics show only moderate sensitivity to perturbed idiom translations, and they fail to reliably detect idiom error spans.

Overall, IDIOMEVAL provides a comprehensive framework for evaluating Chinese idiom translations. The dataset serves as a valuable testbed for investigating various translation systems. Our analysis highlights the inadequacy of existing metrics in evaluating idiom translation and the failures of existing MT systems to produce high-quality idiom translation, which pave the way for future improvement. We will publicly release our taxonomy, code, and annotations.

2 IDIOMEVAL: An Evaluation Framework

In this section, we discuss the details of our framework for idiom translation evaluation. IDIOMEVAL involves two main steps: annotators select the translation span for the idiom within the context and classify it into 9 high-level linguistically grounded categories that are further broken down into 14 subcategories; annotators then rate the severity of the

translation errors and provide their confidence in the annotations. Our annotation interface is built with Thresh (Heineman et al., 2023), and we will release our configuration for public use.

2.1 Select Translation Span and Category

The framework begins with selecting the corresponding translation span of the idiom. Annotators are then asked to choose one category that best describes the idiom translation. For certain categories, annotators are also required to select a subcategory. The IDIOMEVAL categories are as follows:

Good. Translation is of high quality. There are two subcategories under a good translation: 1) The underlying meaning of the idiom is translated well, and 2) The idiom is translated literally but coherent with its surrounding context.

Mistranslation. Translation is incorrect due to wrong choices of words that disrupt idiom understanding.

Unnatural. Translation is understandable yet suboptimal, and it can be improved.

Literal. Translation is literal and coherent with the surrounding context.

Addition. Translation contains irrelevant content beyond the accurate meaning.

Partial. Given idiom is translated partially. There are three subcategories: 1) Translation captures the essence but omits modifiers, 2) Translation captures the essence but with inaccurate modifiers, and 3) Part of the core meaning is missing from the translation.

Repetition. The translation contains repeated correct content or synonyms.

No Translation. Translation lacks the idiom's meaning in the output. There are two subcategories: 1) Idioms are copied or paraphrased in Chinese, and 2) Translation is missing in the output.

More than One. A combination of more than one error category above.

Table 1 displays examples for each translation category. We provide a detailed description of each error type and further examples in Appendix A.

2.2 Rate Severity and Confidence.

As each translation category can affect the overall sentence quality to varying degrees, we ask annotators to rate the severity of the errors. Following (Karpinska et al., 2022; Dou et al., 2022), we define three levels of severity: minor (1), somewhat (2), and a lot (3). Finally, we ask annotators their confidence that their error selection is correct on the same scale. An example annotation for *Partial Translation: Missing Modifier* is shown in Appendix C.1.

3 Data Collection and Annotation

In this section, we describe our methodology for automatically constructing a dataset of 339K Chinese sentences containing idioms across four domains. Using a subset of this data, we then collect and annotate 500 translations from 5 modern systems, which are analyzed in §4 and §5.

3.1 Chinese Idiom Vocabulary

We use the idiom vocabulary provided by Tan and Jiang (2021) for its high coverage, which contains 30,999 idioms with definitions.

3.2 Collect Text with Idioms

To evaluate idiom translations under different contexts, we first collect Chinese texts containing idioms from four domains: News, Web, Wikipedia, and Social Media. To minimize the chance that these texts appear in the pre-training corpora of LLMs used for translation (§3.3), we collect texts that were available after 2023 (referred to as `new corpus`). For each domain, we divide idioms into five groups according to the 1st, 2nd and 3rd quartiles of the overall frequency: VH (very high), H (high), M (medium), L (low), and N (never appear in the data). This allows us to measure the impact of idioms frequency in pre-training corpora. We use the `old corpus` (data before 2023 in the same domains, details in Appendix B.1) to calculate idiom frequencies, as idioms from this period are more likely to be included in the LLMs' training data. Given that idioms tend to have a long-standing presence in a language, their frequency distribution within the same domain is unlikely to change significantly over time. Table 2 displays statistics of the collected data. The details of data collection for each domain are described below:

News. We collect data from Common Crawl News (2023-04), consisting of 26,779 Chinese news articles, which we converted to simplified Chinese for consistency. For each article, we extracted sentences containing idioms, along with the preceding and following sentences for context, resulting in a total of 33,990 instances.

Web. We use Common Crawl (2023-06) to collect web page data. After retrieving the HTML content from each URL, we extracted sentences containing idioms along with their surrounding context,

Domain	Source	Period	Instances	Idioms	Avg. occurrence by freq. range				
					VH	H	M	L	N
News	Common Crawl News	Apr.	33,990	4,783	11.5	3.1	2.2	1.3	1.4
Web	Common Crawl	Jan.-Feb.	233,664	12,108	31.9	2.0	1.3	1.4	7.2
Wikipedia	Wikipedia Meta History	Jan.-Jul.	15,247	4,129	5.3	1.5	1.2	1.1	1.1
Social Media	Weibo Search Results	Jan.-Jul.	56,941	304	475.8	259.8	75.2	20.5	4.1

Table 2: Statistics of collected Chinese data in each domain. *Source* lists the data collection sources. *Period* lists the month of collected data in 2023. *Instances* shows the number of data samples in each domain. *Idiom* column shows the total number of idioms in the collected data. The rightmost column shows the average occurrence of idioms within each frequency range. The number of idioms within each frequency range can be found in Appendix B.2.

similar to the news data. This resulted in a total of 233,664 instances.

Wikipedia. We select the Chinese Wikipedia meta history (2023-07-01) for data collection. We save the revision IDs of articles whose first revision is after the cutoff date. We obtain HTML pages for each revision through Wikimedia REST API[1]. We then extract sentences with idioms and their contexts from the revised text. There are 15,247 instances in total.

Social Media. We use Weibo, a popular Chinese social media platform similar to Twitter, to collect posts containing idioms. We rely on Weibo's search function to collect posts containing idioms. We check the publishing date of each post to ensure it is published after 2023. Unlike the previous domains, where we collected new data with all present idioms from the `old corpus`, querying each idiom will incur a high overhead in API usage. We instead sample 350 idioms from `old corpus` based on frequency range (50 for `VH`, `H`, `M` and 100 for `L` and `N`). We double the size for `L` and `N` to alleviate the scarcity of data. We managed to collect posts for 304 out of 350 idioms, and there are 56,941 instances in total.

3.3 Collect Translations from Systems

We carefully select a diverse set of translation systems to allow us to assess idiom translation performance across a range of system types: GPT-4 (a SOTA LLM) (Achiam et al., 2023), Alma-13B (a fine-tuned translation model) (Xu et al., 2024), Qwen-14B-Chat (Bai et al., 2023) (a bilingual LLM for Chinese and English), mT0-13B (a multilingual LLM) (Muennighoff et al., 2022), and Google Translate (a widely used commercial service).

We sample 5 instances from each frequency range, resulting in 25 Chinese texts to be translated for each domain. In total, this produces 500 translation pairs from the 5 systems. The authors, who are native speakers of both Chinese and English, manually translated each Chinese text into English. These human translations serve as reference translations for our later analysis.

3.4 Human Annotation

We recruit native Chinese speakers fluent in English through Prolific. The annotation process has two phases: the pilot phase and the main phase. In the pilot phase, 20 participants are provided with annotation guidelines and a quiz. The top 5 scorers are selected for the main phase. In the main phase, each translation pair is assigned to three annotators, ensuring equal workload distribution. The final annotations are determined by a majority vote, with a manual resolution by the first author for any ties. More details on inter-annotator agreement and quality control are provided in Appendix C.2.

4 Translation System Evaluation

We conduct a detailed analysis of the idiom translation quality of the modern systems. We discuss how each system performs in each domain and their relative comparison. We then examine how domain and idiom frequency impact translation quality, followed by a discussion on the severity of translation errors. We present our results in Figures 2 to 4. Figure 2 displays the ratio of different translation categories from each system in the four domains. Figure 3 shows the top three most commonly seen errors for each system. Figure 4 illustrates the severity composition from each system. Further analysis such as subcategories selected, error distribution by frequency range, and average severity scores for each category, can be found in Figures 7 to 9 in the Appendix. The following are the key findings:

1. GPT-4 and Qwen have overall higher translation quality, while Alma and mT0 have per-

[1] https://zh.wikipedia.org/api/rest_v1/

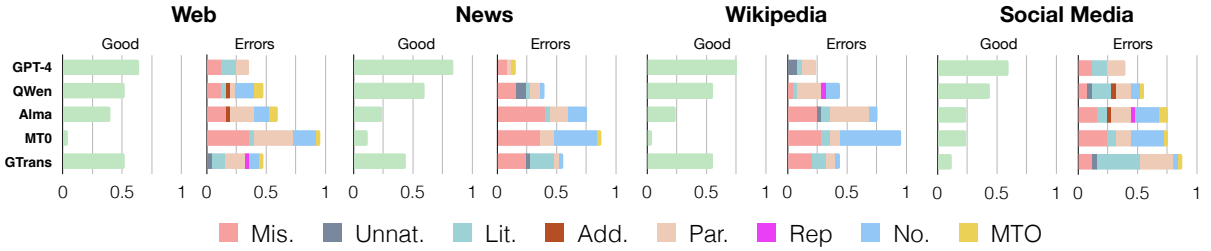

Figure 2: System-centric view of translation results in each domain. Each dataset is accompanied by two subfigures. Left subfigure: ratio of Good Translations. Right subfigure: composition of error categories for each system.

formed consistently badly (Figure 2). GPT-4 has the highest number of Good Translations across all the domains, while Qwen achieves comparable second performance. On `News` and `Wikipedia`, GPT-4 surpasses the second-best system (Qwen) by a margin of 0.24 and 0.2. Despite the best performance among the models, GPT-4 still fails to achieve satisfying results, with more than 25% of all translations being problematic. For Alma and mT0, fewer than 25% of their translations are rated as good (except for Alma on `Web`). This disparity is likely due to a lack of idiom-related Chinese data in the pre-training and fine-tuning phases. For example, Alma's backbone, LLaMA-2, is primarily pre-trained on English, limiting its performance. In contrast, Qwen, which is pre-trained on both Chinese and English bilingual data, achieves better translation quality.

2. Systems' translation quality varies by domain, where they perform best on `News` but worst on `Social Media` (Figure 2). Although GPT-4 and Qwen provide the highest translation quality overall, their number of Good Translations varies significantly across domains. For GPT-4, 84% of its translations are rated as good on `News`, but this drops to 60% on `Social Media`. Qwen shows similar fluctuations. Overall, translation quality is worse on `Social Media`. This may be because social media often features newly emerging contexts, where idioms are used in novel ways compared to other domains. These variations highlight intrinsic differences in idiom usage across domains and emphasize the need for diverse training data to ensure consistently high-quality translations.

3. Frequency ranges of idioms are not correlated with translation quality (Figures 8 and 9). For idioms that appear in `old corpus`, their translation quality does not decrease with their frequency ranges. For instance, idioms in `Web` from the M range (100% Good Translations) get the best translation by GPT-4, while those from the H and L range (60% Good Translations) achieve comparable translation quality. Similar patterns can be found in other models. Interestingly, idioms that never appeared in the `old corpus` (N range) do not always result in poor translations. For instance, on `Wikipedia`, Alma produces the highest-quality translations for N-range idioms over other frequencies. This observation is particularly remarkable on Social Media data, where idioms that never appear get the best translation quality by almost all models. Since the N range is defined regarding the appearance of idioms in `old corpus` for that domain. Idioms not showing up in one domain can still appear in another domain and be included in the pre-training corpus. This shows that language models transfer the knowledge from a different domain and translate the idiom correctly.

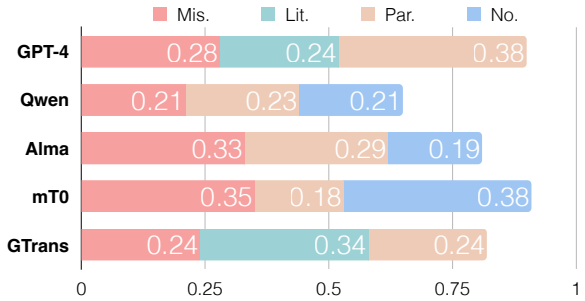

Figure 3: Top three most commonly seen errors for each system. `Mis.`: Mistranslation; `Lit.`: Literal Translation; `Par.`: Partial Translation; `No.`: No Translation.

4. Translation error types differ across systems, but Mistranslation and Partial Translation are commonly seen (Figures 2, 3, 8 and 9). GPT-4 exhibits the most Mistranslation and Partial Translations on `Web` and `Social Media`. Surprisingly, it has no Mistranslation on `Wikipedia`. Instead, 20% of its translations in the H and N range are marked as unnatural. It also suffers from more than one error on `News` (20% in the M range). Qwen and Alma produce more diverse errors, with at

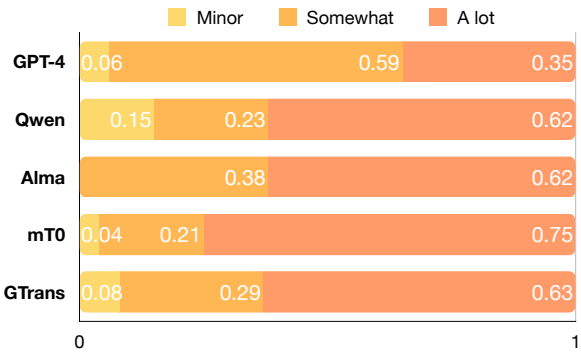

Figure 4: Ratio of each severity score for each system, calculated overall all errors and domains.

least three distinct error types across several frequency ranges and domains. They are the only two models producing Addition errors (`Web` and `Social Media`). Alma's Partial Translations vary by domain: it omits modifiers on `News` (100%) and `Wikipedia` (75%), but misses core meanings on `Web` (100%) and `Social Media` (67%). mT0 frequently produces No Translations, particularly on `Wikipedia`, where it fails to translate at least 40% of idioms across frequency ranges. Google Translate suffers from Literal Translations, with 36% of its translations marked as literal on four frequency ranges from `Social Media`. It also produces Unnatural Translations across three domains (4%).

5. Systems produce errors with different severity, with mT0 producing the most severe errors and GPT-4 making less severe ones. Errors that interrupt meaning are more severe than errors that preserve the overall meaning (Figures 4, 7 and 8). When models make errors, over half of them significantly impact the understanding of the text. Despite GPT-4's top performance, only 6% of its incorrect translations have a minor impact, placing it third behind Qwen and Google Translate. Alma makes the most severe errors, with none of its errors classified as minor.

Mistranslation and No Translation are the most severe error types, as they disrupt meaning significantly. Mistranslation has an average severity score of 2.43 and No Translation score of 2.8. In contrast, Unnatural Translation (max severity of 2), Addition (max severity of 2.5), and Repetition (max severity of 2) are less severe, as they typically preserve some correct or near-correct idiom parts.

5 Automatic Metric Evaluation

We evaluate how well current automatic translation metrics assess idiom translations, focusing on their ability to evaluate translations containing idioms, their sensitivity to idiom translation errors, and whether they can detect specific idiom translation error spans. Additionally, we experiment with the recent GPT-4o (OpenAI, 2024) to see if it could accurately identify idiom translation errors.

5.1 Metric Performance

We consider six of the most widely-used and strongest automatic metrics (Freitag et al., 2023): BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019), COMET (Rei et al., 2022a), COMETKIWI (Rei et al., 2022b), MetricX (Juraska et al., 2023) and MetricX-QE (Juraska et al., 2023). Among these, COMETKIWI and MetricX-QE are reference-free. To allow for comparison with human annotations, we collapse each annotation into a single numeric score, which is the severity score (from 1 to 3) directly as the overall score. Good translations are marked as 0.

To understand how well these metrics evaluate idioms both in context and individually, we calculate their scores based on the full context and only the idiom part. Table 3 reports the Pearson correlation coefficient (r) between the metrics and human evaluation for different error types. We found that, at both the full context and idiom levels, the metrics struggle to provide effective fine-grained measurements, with the highest correlation being 0.38 for full context and 0.52 for idioms. In most cases, correlations are stronger when measured on idiom translations than in full text, as the surrounding context of idioms also influences the metric scores. Nevertheless, existing metrics show weak correlations with human ratings, thus failing to measure the quality of idiom translation from a human-like perspective. Similar trends are observed when examining the ranking correlation with Kendall's τ (details in Appendix E.1).

Given that the metrics cannot effectively measure the severity of errors, we next investigate a simpler question: Can they distinguish good idiom translations from bad ones? We report the ROC-AUC score of each metric and the percentage of pairs where good translations score better than bad ones in Table 4 BERTScore and COMET perform the best on both metrics. However, in more than 20% cases, good translations are still rated lower than those with errors, indicating that these metrics are not consistently reliable for distinguishing between good and bad idiom translations.

Scope	Category	Size	Reference-Based Metrics				Reference-Free Metrics	
			BLEU	BERTScore	COMET	MetricX	COMETKIWI	MetricX-QE
Full	Mistranslation	87	-0.029	0.089	0.102	0.070	0.180	0.017
	Literal	44	0.012	0.149	0.251	0.387	0.170	0.236
	Partial	73	-0.008	0.051	0.139	0.159	0.171	0.147
	No Translation	63	-0.045	0.149	0.150	0.008	0.150	0.043
Idiom	Mistranslation	87	-0.106	0.072	0.234	0.143	0.097	0.147
	Literal	44	0.231	0.264	0.368	0.376	-0.219	-0.085
	Partial	73	0.163	0.008	0.194	0.175	0.170	0.044
	No Translation	63	*	*	0.185	-0.117	0.200	0.207
Full	All errors	296	0.087	0.253	0.275	0.188	0.273	0.148
	All categories	500	0.252	0.387	0.380	0.229	-0.333	-0.239
Idiom	All errors	296	0.282	0.307	0.454	0.248	0.062	0.177
	All categories	500	0.364	0.431	0.524	0.417	0.154	0.286

Table 3: Pearson's r between metrics and human annotations. Evaluation is measured on both *Full* translation and *Idiom* translation. More Than One is omitted here as all its instances are rated as the highest severity. Unnatural, Addition and Repetition are omitted due to small sample size. *: for No Translation on *Idiom*, BLEU and BERTScore outputs 0 and thus omitted. We also present Kendall's τ is in Table 9 in Appendix E.1, which shows similar patterns.

Metric	ROC-AUC		Good > Bad	
	Full	Idiom	Full	Idiom
BLEU	0.64	0.70	75%	56%
BERTScore	0.70	0.74	79%	74%
COMET	0.70	0.77	80%	79%
MetricX	0.61	0.74	70%	79%
COMETKIWI	0.66	0.58	68%	58%
MetricX-QE	0.61	0.66	69%	65%

Table 4: Distinguishing good and bad translations based on ROC-AUC scores and ranking of translations using metrics output on *Full* text and *Idiom* translations.

5.2 Metric Sensitivity

We measure metric sensitivity based on a similar approach to Karpinska et al. (2022) by comparing how each metric scores gold translations versus perturbed translations. Metrics that are sensitive to perturbations can accurately identify the perturbations. We focus on reference-free metrics, COMETKIWI and MetricX-QE as reference-based metrics use the gold translation as reference. To obtain perturbed translations, we use GPT-4o (OpenAI, 2024) to edit the existing reference translations into different errors. The task is decomposed into two steps: 1) Extraction: extracting the idiom translation from the English sentence, and 2) Perturbation: perturbing idiom translations with errors (see Appendix E.2 for the prompts). For each error type, we randomly sampled 10 instances and checked if the edited translations matched the intended error category. More than 70% of the instances were edited correctly for each category. Since the model was specifically instructed to introduce errors, the gold translation should always be better than the perturbed ones.

Perturbation	COMETKIWI		MetricX-QE	
	Full	Idiom	Full	Idiom
Mistranslation	0.65	0.57	0.69	0.73
Unnatural	0.48	0.49	0.62	0.59
Literal	0.59	0.50	0.76	0.60
Addition	0.62	0.27	0.47	0.58
Partial	0.56	0.53	0.62	0.55
Repetition	0.63	0.52	0.77	0.52
No Translation	0.71	0.84	0.83	0.46

Table 5: Sensitivity of reference-free metrics. We measure sensitivity by the ratio of instances where unperturbed translations score better than the perturbed translations. *Full* column represents the sensitivity measured on the full text while *Idiom* column measures idiom translation only.

We calculate the accuracy on perturbation as the ratio of instances where gold translations score better than perturbed translations, as shown in Table 5. For full-context translations, the top-performing metric, MetricX-QE, achieves less than 0.7 accuracy in most categories, with the highest being 0.83 for No Translation category. However, when evaluating translations on the idiom part, the metrics show even lower performance, showing they are less effective at detecting errors when only the idiom is presented. Overall, these results further imply using these metrics on idiom translation tasks could lead to inaccurate conclusions.

	P	R	F_1	WP	WR	WF_1
xCOMET-XL	12.5	44.7	16.6	0.4	2.2	0.7
xCOMET-XXL	10.2	30.3	12.4	0.5	1.6	0.6
xCOMET-XL (src)	13.3	49.7	18.1	0.5	2.4	0.7
xCOMET-XXL (src)	9.3	11.4	30.0	0.5	1.6	0.5

Table 6: Precision and F_1 scores for error span detection using xCOMET. *src* means reference-free version. *W* represents weighted metric output. All values are multiplied by 100 for readability.

5.3 Can Existing Metrics Identify Idiom Error Span?

We use xCOMET (Guerreiro et al., 2023), the state-of-the-art error span detection model, to examine if it can accurately detect the span of idiom translation errors. We evaluate the performance using character level F_1 score and only consider target spans. Similar to Blain et al. (2023), we also compute weighted F_1 scores by converting severity scores into binary weights.

Table 6 displays the corresponding performance. We find that xCOMET performs poorly in detecting the span of idiom translation error, with the best model achieving only 0.3 F_1.

5.4 Can GPT-4o Identify Errors?

We additionally inspect GPT-4o's ability to identify idiom translation categories. We decompose this task into two steps: 1) Extraction (see §5.2), and 2) Identification: identifying the categories of the extracted idiom translation (see Appendix E.2 for prompts). We compare identified categories with human-annotated categories and find that GPT-4o achieves an accuracy of 42%, highlighting its limitations in accurately evaluating idiom translations.

6 Related Work

Machine Translation with Idioms. We consider multi-word expression (MWE) a close line of work to idioms, and there have been plenty of studies on MWEs (Sag et al., 2002; Calzolari et al., 2002). Zaninello and Birch (2020) explored annotation and data augmentation techniques for MWE translations. Fadaee et al. (2018) add additional input features by pretending special tokens to indicate the presence of idioms. Gamallo and Garcia (2019) proposed compositional translation using cross-lingual embeddings. More recent works focus on idiom identification (Tedeschi et al., 2022), characterization (Socolof et al., 2021) and representation (Zeng and Bhat, 2022). Dankers et al. (2022b); Baziotis et al. (2023) provide an analysis of idiom processing with Transformers in the context of neural machine translation. Liu et al. (2023) provide a characterization of idiomatic translation and its issues and introduce upweighting and a retrieval module to improve the translation. In contrast to these works, our research introduces a fine-grained evaluation framework for idiom translation and highlights the limitations of modern LLMs, such as GPT-4, and automatic metrics in generating and evaluating idiom translations.

Studies with Chinese Idioms. There has been limited research involving Chinese idioms. Zheng et al. (2019) proposed a large-scale Chinese idiom dataset ChID for cloze tests. It is the first work to study Chinese idiom understanding in the form of machine reading comprehension. Tan and Jiang (2021) constructs a dataset to facilitate the evaluation of Chinese idiom embeddings. Tan et al. (2021) collected a large pretaining corpus of Chinese text by crawling e-books online. Liao and Cheng (2023) studied text polishing with Chinese idioms, where sentences are rewritten to be more elegant with idioms. Other works have focused on creating parallel corpus with Chinese idioms. Shao et al. (2018) constructed the CIBB dataset that contains the English translations of 50 distinct Chinese idioms, together with a blacklist of literal translations. Wang and Yu (2010) created CIKB dataset that contains about 38k idioms and different English translations. However, only 28% of all samples have complete translations, and the dataset is not publicly available. Tang (2022) constructed PETCI dataset that contains idioms collected from a dictionary, together with translation results from Google and DeepL.

7 Conclusion

We introduce IDIOMEVAL, a framework with a carefully designed taxonomy that facilitates evaluating Chinese idiom translation. We posit that our framework can be extended to different languages. Our evaluation of various systems across multiple domains reveals that they struggle to produce high-quality idiom translations, with performance varying by domain and showing no correlation with idiom frequency. Additionally, we find that existing metrics fail to correlate with human ratings or reliably distinguish good translations from poor ones. Our work lays the groundwork for future research to improve idiom translation performance and to develop metrics focused on idioms.

Limitations

First, IDIOMEVAL only includes snapshots of data collected from four domains. We choose not to include more snapshots or domains due to the time cost. However, we believe our observations remain consistent with more data. Second, we only annotated 500 translation pairs on Prolific due to time and money constraints. Despite limitations, these annotations still provide valuable insights into how LLMs and commercial system perform on the Chinese idiom translation task. Third, we only test five translation models. Given the rapid evolution of LLMs, it is necessary to examine the performance of newly released models. However, our aim is not to test all the available LLMs but rather to develop a framework for evaluating Chinese idiom translation. We leave the evaluation of more models to future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M Guerreiro, Diptesh Kanojia, José GC de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, et al. 2023. Findings of the wmt 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine Macleod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *International Conference on Language Resources and Evaluation*.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022a. Can transformer be too compositional? analysing idiom processing in neural machine translation. *ArXiv*, abs/2205.15301.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022b. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. *ArXiv*, abs/1802.04681.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, et al. 2023. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

Pablo Gamallo and Marcos Garcia. 2019. Unsupervised compositional translation of multiword expressions. In *MWE-WN@ACL*.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

David Heineman, Yao Dou, and Wei Xu. 2023. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–345.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kenneth Katzner and Kirk Miller. 2002. *The languages of the world*. Routledge.

Junwei Liao and Shuai Cheng. 2023. Text polishing with chinese idiom: Task, datasets and pre-trained baselines. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

OpenAI. 2024. Hello gpt-4o.

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.

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. 2018. Evaluating machine translation performance on Chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Michaela Socolof, Jackie Chi Kit Cheung, Michael Wagner, and Timothy J. O'Donnell. 2021. Characterizing idioms: Conventionality and contingency. *ArXiv*, abs/2104.08664.

Minghuan Tan and Jing Jiang. 2021. Learning and evaluating chinese idiom embeddings. In *Recent Advances in Natural Language Processing*.

Minghuan Tan, Jing Jiang, and Bingtian Dai. 2021. A bert-based two-stage model for chinese chengyu recommendation. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20:92:1–92:18.

Kenan Tang. 2022. Petci: A parallel english translation dataset of chinese idioms. *ArXiv*, abs/2202.09509.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *NAACL-HLT*.

Lei Wang and Shiwen Yu. 2010. Construction of Chinese idiom knowledge-base and its applications. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 11–18, Beijing, China. Coling 2010 Organizing Committee.

Ran Wang. 2022. The cultural differences between english and chinese idioms and translation. In *2022 International Conference on Creative Industry and Knowledge Economy (CIKE 2022)*, pages 610–614. Atlantis Press.

You Wang. 2021. A contrastive research on the definitions and categories of chinese and english idioms. In *7th International Conference on Humanities and Social Science Research (ICHSSR 2021)*, pages 423–426. Atlantis Press.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *ICLR*.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *International Conference on Language Resources and Evaluation*.

Ziheng Zeng and S. Bhat. 2022. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

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

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. Chid: A large-scale chinese idiom dataset for cloze test. *ArXiv*, abs/1906.01265.

A Idiom Translation Examples

Below are the input and output of different translation categories and subcategories, along with a corresponding explanation. We skip trivial cases.

Good: Correct Underlying Meaning
Chinese: 主断层绕过核电站，停了下来。美国地质调查局的科学家们欢呼雀跃，却发现断层并未被完全阻止。
English: The main fault bypassed the nuclear power plant and came to a stop. Scientists from the United States Geological Survey cheered and rejoiced, but they discovered that the fault had not been completely halted.
Explanation: 欢呼雀跃 is translated correctly into "cheered and rejoiced".

Good: Literal and Coherent
Chinese: 此举指向性明确——挑衅美元，直接让联储局骑虎难下。（更多内容见：《沙特为首的油霸们再捅刀美元项庄舞剑意在沛公》）3月，沙特一连串的'出牌'，是沙特战略转型的重要一步。
English: This move has a clear direction - provoking the US dollar and putting the Federal Reserve in a difficult position. (For more content, see: "Saudi-led oil tycoons stab the US dollar again, Xiang Zhuang's sword dance is aimed at Pei Gong") In March, Saudi Arabia's series of 'moves' is an important step in its strategic transformation.
Explanation: The original meaning of 项庄舞剑 means the actual intention is not what it appears to be. However, in this case, the idiom is used to refer to the historical story and a literal translation is more appropriate.

Mistranslation
Chinese: 另一半有时需要一个人安静地思考，此时你不必过多打扰，问得太多只会自寻烦恼。
English: The other half sometimes needs someone to think quietly by themselves, at which point you don't need to disturb them too much. Asking too many questions will only bring them problems.
Explanation: 自寻烦恼 means asking for trouble or trouble someone him/herself, not "bringing other people trouble".

Unnatural
Chinese: 善战者无赫赫之功。
English: A good warrior does not achieve a glaring victory.
Explanation: 赫赫之功 means "a great achievement" or "a great victory". "A glaring victory" conveys a rough meaning but the usage of "glaring" is not perfect enough.

Literal Translation
Chinese: 我爸临终时候说，这一生身无长处，不曾留下什么给我
English: My father said when he was dying, 'I have no merits in this life, I have not left anything for you.'
Explanation: 身无长处 is translated literally here. 身: in this life. 无: no. 长处: merits.

Addition
Chinese: 程长庚不仅"于剧学素有深造，无所不通"，而且怀有崇高的戏剧理想和抱负。
English: Cheng was not only "well-versed in drama studies, knowledgeable and passionate in all aspects", but also harbored lofty theatrical ideals and ambitions.
Explanation: 无所不通 means knowledgeable in all aspects, and it is not related to one's passion.

Partial Translation: Missing Modifier
Chinese: 《易》说："出征顺利，斩了首恶，俘获他的同伙。"称扬赞美诛杀罪大恶极之人，那些不归顺的都来归服了。
English: The "I Ching" says, "The expedition is successful, the chief evil is beheaded, and his accomplices are captured." It praises the execution of those who commit evil crimes, and those who do not submit have come to surrender.
Explanation: 罪大恶极 refers to extremely evil crimes. Here, the translation has conveyed the core idea of the idea, but the degree of "evil" is missing.

Partial Translation: Missing Core Info
Chinese: 再者，平台繁多，此处不留人，自有留人处，对于这类无良民宿，仅靠平台按现行规则管制，并不足以纠偏。
English: Also, there are many agencies, there is no place to stay, and against these bad inns, only the agencies with the current rules cannot adequately curb the abuse.
Explanation: Only the first half of 此处不留人，自有留人处 is translated. The second half of the idiom contains important information for us to understand the idiom, too.

Partial Translation: Inaccurate Modifier
Chinese: 被俘俄军惊恐万状，其实他比地上的同伙幸运多了。

Domain	Source	Period	Instances	Idioms	Avg. occurrence by freq. range			
					VH	H	M	L
News	News Crawl	2019	17,931,873	10,830	425.6	38.6	8.0	1.5
Web	mc4	2008-22	54,542,308	30,652	8135.5	220.0	53.0	26.0
Wikipedia	wiki2019zh	2018-19	1,043,224	11,754	95.1	13.0	3.4	1.0
Social Media	PchatbotW	2008-22	139,448,339	15,089	2313.1	189.9	16.2	1.8

Table 7: Statistics of *old corpus*. *Source* lists the data collection sources. *Period* lists the year of collected data. *Instances* shows the number of data samples in each domain. *Idiom* column shows the total number of idioms. The rightmost column shows the average occurrence of idioms within each frequency range.

English: The captured Russian soldiers are a bit terrified, but in reality, they are much luckier than their comrades on the ground.
Explanation: 惊恐万状 means "very terrified". The translation inaccurately used "a bit" as a modifier.

Repetition
Chinese: 我认为最佳的父子关系应该如兄如弟。
English: I think the best relationship between a father and son should be like that of a brother and a brother.
Explanation: 如兄如弟 means the relationship is like brothers. In this case, the translation is repeated.

No Translation: Missing from Output
Chinese: 伊拉斯谟当时正从布隆泉飞向普利托利亚，除了不速之客的蛇外，机上还载了4名乘客。
English: Irascible was flying from Bloemfontein to Pretoria and had four other passengers on board, including the snake.
Explanation: 不速之客 is not translated and is missing from the English sentences.

No Translation: Chinese Idioms Copied or Paraphrased
Chinese: 他指平日儿子和媳妇都早出晚归，只有他们两老和两名孙女在家，他担心对方会变本加厉。
English: He指平日儿子和媳妇都早出晚归，只有他们两老和两名孙女在家，他担心对方会变本加厉。
Explanation: 早出晚归 is not translated and remains in Chinese text.

B Statistics of Corpus

B.1 Statistics of Old Corpus

Table 7 shows the statistics of the old corpus.

B.2 Number of Idioms within Each Frequency Range

Domain	VH	H	M	L	N
News	2,402	1,548	554	162	117
Web	7,040	2,985	1,195	814	74
Wikipedia	2,453	998	390	121	167
Social Media	50	50	50	94	60

Table 8: Number of idioms within each frequency range in our collected data.

Table 8 displays the number of idioms in each frequency range in our newly collected data.

C Annotation Details

C.1 Annotation Interface

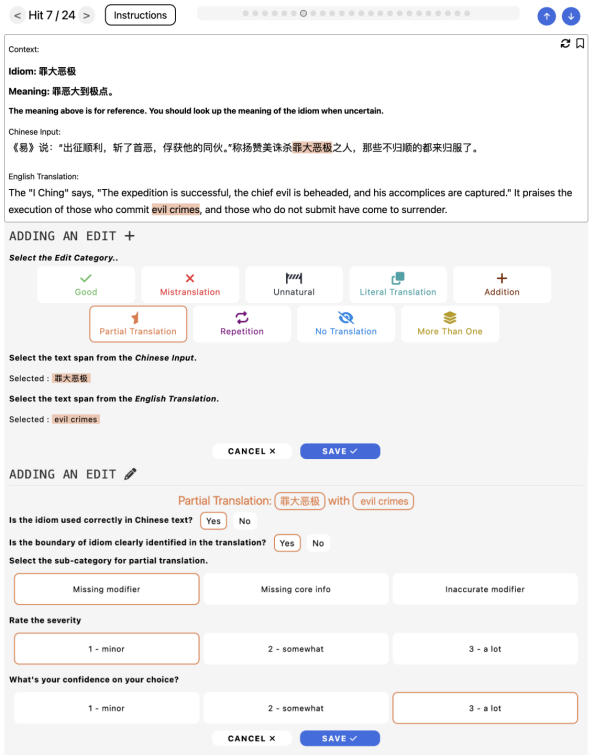

Figure 5: Idiom annotation framework.

Figure 5 shows IDIOMEVALannotation interface designed using Thresh (Heineman et al., 2023).

C.2 Quality Control

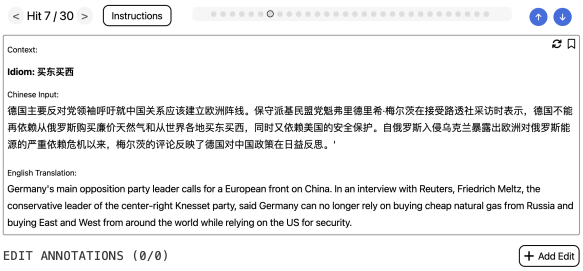

Figure 6: Example quiz question.

We use Prolific to recruit annotators who are native Chinese speakers fluent in English. We only collect the annotations on the data we provide, and no information about the annotators themselves is collected. We have received full consent from the annotators on this task. The annotation process is divided into two phases: *pilot phase* and *main phase*. In the pilot phase, 20 participants are selected for pre-screening. They are provided with detailed guidelines and must complete a quiz to demonstrate their understanding. The quiz consists of 30 questions, covering each category and subcategory with 2-3 questions. An example quiz question is shown in Figure 6. The top 5 performers from the pilot phase are selected for the main phase. In the main phase, we randomly assign each translation pair to three annotators while we ensure each annotator sees an equal number of pairs. We sample 100 annotations and evaluate them against the authors' own annotations, achieving a category accuracy of 66% and subcategory accuracy of 88%. For the final annotations, we take the majority vote and the first author manually resolves any tie cases. This resolution improves the average Cohen's κ among annotators on category (subcategory) agreement from 0.47 (0.38) to 0.73 (0.69), indicating strong annotator agreement given some translations may have more than one answer. For instance, translating 虚情假意 (false feelings and intentions, indicating insincerity in one's intention) into "false" can be considered as either a Mistranslation or a Partial Translation. Annotators are compensated at $22 per hour.

D Further Analysis on Translation Systems

D.1 Severity Scores

Figure 7 shows the breakdown of the severity score of each system on each error category across domains. We observe that systems have severity scores above 2 in all domains. Idiom translation errors that interrupt understanding of the sentence generally have higher severity scores.

D.2 Category Breakdown

Figure 8 shows the subcategory composition for Good translation, Mistranslation and Partial Translation. It can be seen that subcategories vary across domains and systems. For instance, for Partial Translations, On `News` and `Wikipedia`, Alma is likely to omit modifiers (100% and 75%), while on `Web` and `Social Media`, it misses the core meaning in most cases (100% and 67%). mT0 is accomplished by a significant number of No Translations in each domain, where its translations are mostly missing from the output.

D.3 Frequency Breakdown

Figure 9 illustrates each system's performance in each frequency range across domains. Idioms from lower frequency ranges and idioms that never appear in the `Old Corpus` (thus probably also in the pre-training data of systems) are not necessarily translated worse than those commonly seen. It can be seen that frequency ranges of idioms are not correlated with translation quality.

E Further Analysis on Evaluation Metrics

E.1 Kendall's τ between Evaluation Metrics and Human Annotations

Table 9 shows Kendall's τ between evaluation metrics and human annotations. Similar to Table 3, Kendall's τ remains weak across categories, indicating low agreement between ranking based on metric scores and human ratings.

E.2 Prompts

Prompt 1 shows prompts for extracting idiom translations. Prompt 2 shows prompts for editing idiom translations. Prompt 3 shows prompts for identifying idiom translations.

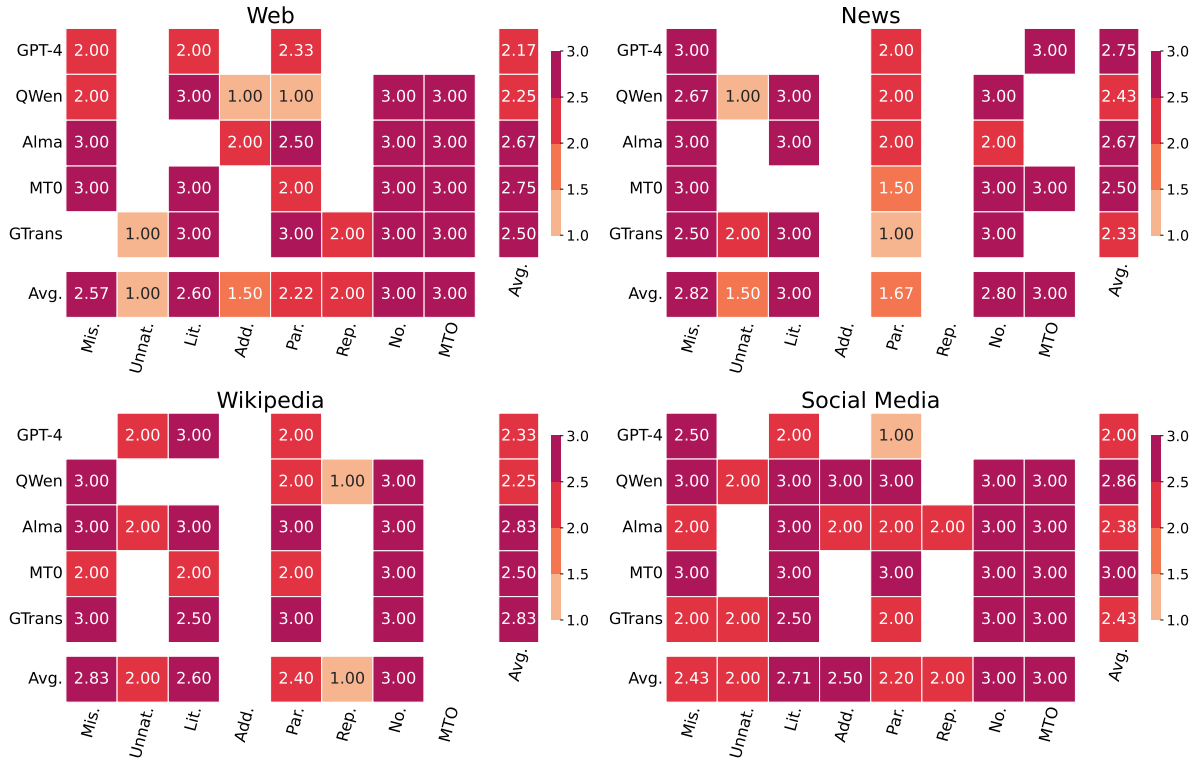

Figure 7: Average severity scores for each translation category in each domain. The rightmost column represents the average severity score across all errors for each system. The bottom row represents the average severity score across all systems for each error.

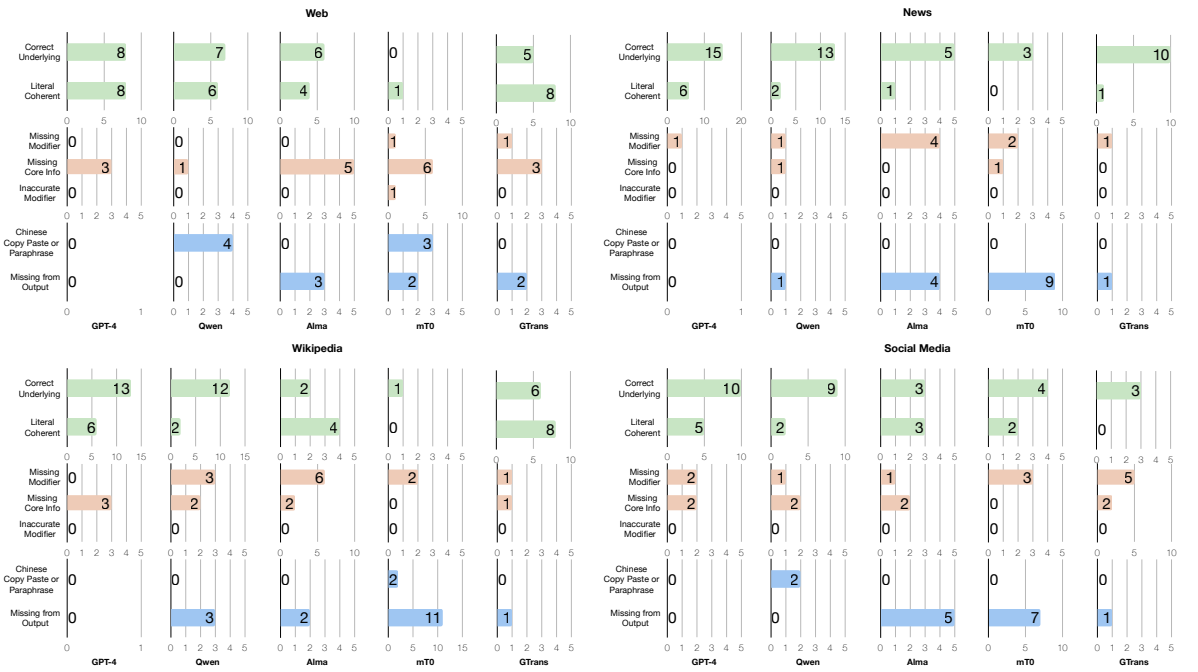

Figure 8: Subcategory composition for Good Translation, Partial Translation and No Translation.

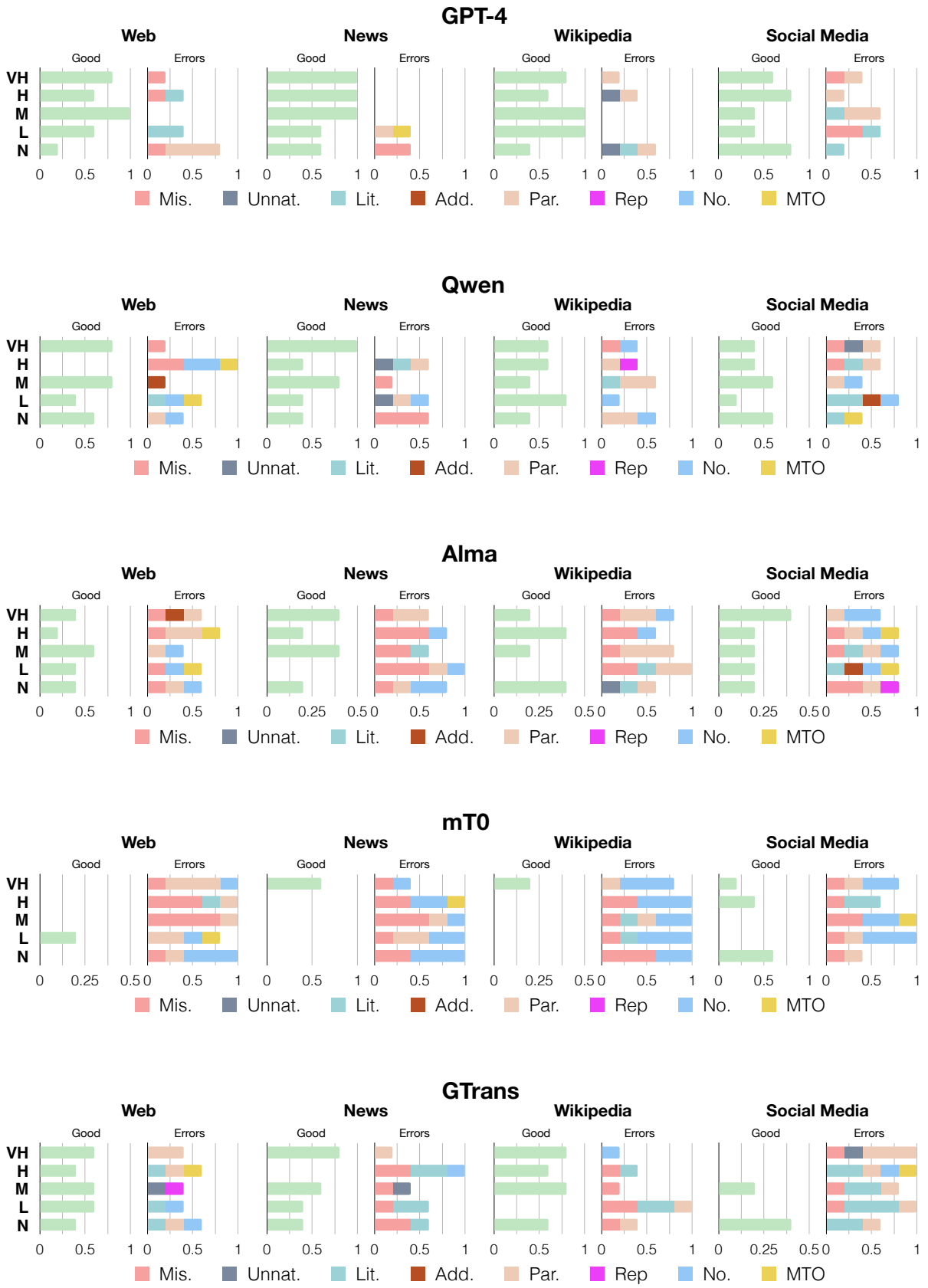

Figure 9: Frequency-centric view of translation results in each domain. There are two subfigures in each dataset. Left subfigure: mean ratio of good translations across all systems. Right subfigure: mean ratio of error categories across all systems.

Scope	Category	Size	Reference-Based Metrics				Reference-Free Metrics	
			BLEU	BERTScore	COMET	MetricX	COMETKIWI	MetricX-QE
Full	Mistranslation	87	0.021	0.072	0.086	0.076	0.173	0.005
	Literal	44	-0.005	0.101	0.210	0.335	0.154	0.157
	Partial	73	-0.029	0.014	0.031	0.154	0.122	0.125
	No Translation	63	0.034	0.143	0.116	0.016	0.153	0.093
Idiom	Mistranslation	87	-0.030	0.013	0.191	0.154	0.119	0.099
	Literal	44	0.123	0.179	0.295	0.295	-0.164	-0.192
	Partial	73	0.085	-0.004	0.141	0.147	0.124	0.021
	No Translation	63	*	*	0.161	-0.032	0.175	0.220
Full	All errors	296	0.077	0.200	0.216	0.175	0.235	0.123
	All categories	500	0.195	0.303	0.306	0.186	0.267	0.172
Idiom	All errors	296	0.223	0.270	0.343	0.264	0.037	0.161
	All categories	500	0.339	0.371	0.423	0.373	0.110	0.239

Table 9: Kendall's τ between evaluation metrics and human annotations. Evaluation is measured on both *Full* translation and *Idiom* translation. Good is omitted since it does not have severity scores. More Than One is omitted here as all its instances are rated as the highest severity. Unnatural, Addition and Repetition are omitted due to small sample size. *: for No Translation on *Idiom*, BLEU and BERTScore outputs 0 and thus omitted. Existing metrics fail to produce measures that are correlated strongly with human annotations.

Prompt 1: Prompt for extracting idiom translations

```
## Task
Analyze the provided Chinese
   idiom, Chinese sentence
   containing the idiom and its
   English translation to extract
    corresponding idiom
   translation.

## Input to Analyze
* Chinese idiom: [PLACEHOLDER]
* Chinese sentence: [PLACEHOLDER]
* English translation: [
   PLACEHOLDER]

Please note:
1. Ensure the extracted idiom
   translation is short and
   concise, and do not include
   irrelevant translation.
2. If the given idiom appears
   multiple times in the Chinese
   sentence, only analyze the
   first occurrence.
3. If no corresponding
   translation is found, output
   an empty string.

Read the input carefully. Write
   down a brief thought process
   first, with the 3 notes in
   mind. Then extract the
   translation of the idiom.
Respond in a JSON format {"Idiom
   translation": }. The key must
   be "Idiom translation", and
   the value must be the
   corresponding idiom
   translation in the English
   sentence.
```

Prompt 2: Prompt for perturbing idiom translations

```
## Task
Analyze the provided Chinese
   idiom, idiom meaning, Chinese
   sentence containing the idiom
   and its English translation.
The goal is to edit the given
   idiom translation to comply
   with the given category.


## Categories
There are 7 categories to
   consider:
* Mistranslation: idiom
   translation is incorrect due
   to wrong choices of words or
   phrases, and it affects our
   understanding of the
   translated sentence.
* Partial Translation: idiom is
   translated partially. Part of
   the idiom meaning is missing
   or the extremity is inaccurate
   .
* Repetition: translation of the
   idiom contains repeated words
   or phrases, or their synonyms.
* Unnatural: the translation is
   ok but not perfect enough due
   to improper choices of words
   or grammar errors. In other
   words, the translation can be
   improved by more appropriate
   choices of words.
* Literal Translation: idiom is
   translated literally and
   translation is not coherent
   with context.
* Addition: In addition to good
   translation of idiom, the
   translation also contains non-
   present content in the source.
* No Translation: there is no
   translation of the idiom in
   the output.

## Input to Edit
* Chinese idiom: [PLACEHOLDER]
* Idiom meaning: [PLACEHOLDER]
* Chinese sentence: [PLACEHOLDER]
* English translation: [
   PLACEHOLDER]
* Idiom translation: [PLACEHOLDER
   ]

Please note:
1. Given category only applies to
    the idiom translation itself.
2. An idiom can have multiple
   meanings. Find the most
   appropriate meaning under the
   given context.


Read the input carefully. YOU
   NEED TO MAKE THE IDIOM
   TRANSLATION A [PLACEHOLDER].
First, modify the idiom
   translation to comply with the
    given category.
Then, rewrite the given
   translation by replacing the
   original idiom translation
   with the new translation.

Do not include your thought
   process.

Output the corresponding edited
   translation only. Do not
   include anything else in your
   output.
```

Prompt 3: Prompt for editting idiom translations

```
## Task
Analyze the provided Chinese
   idiom, idiom meaning, Chinese
   sentence containing the idiom
   and its English translation to
    identify idiom translation
   category.

## Categories
There are 8 categories to
   consider:
* Good Translation: the idiom is
   translated perfectly.
* Mistranslation: idiom
   translation is incorrect due
   to wrong choices of words or
   phrases, and it affects our
   understanding of the
   translated sentence.
* Partial Translation: idiom is
   translated partially. Part of
   the idiom meaning is missing
   or the extremity is inaccurate
   .
* Repetition: translation of the
   idiom contains repeated words
   or phrases, or their synonyms.
* Unnatural: the translation is
   ok but not perfect enough due
   to improper choices of words
   or grammar errors. In other
   words, the translation can be
   improved by more appropriate
   choices of words.
* Literal Translation: idiom is
   translated literally and
   translation is not coherent
   with context.
* Addition: In addition to good
   translation of idiom, the
   translation also contains non-
   present content in the source.
* No Translation: there is no
   translation of the idiom in
   the output.

## Input to Analyze
* Chinese idiom: [PLACEHOLDER]
* Idiom meaning: [PLACEHOLDER]
* Chinese sentence: [PLACEHOLDER]
* English translation: [
   PLACEHOLDER]
* Idiom translation: [PLACEHOLDER
   ]

Please note:
1. Focus on analyzing the
   translation of the idiom under
    the context of the sentence.
2. Compare the idiom meaning and
   provided idiom translation to
   make the judgement. Note an
   idiom can have multiple
   meanings.
3. If the given idiom appears
   multiple times in the Chinese
   sentence, only analyze the
   first occurrence.
4. If the idiom translation only
   captures the essence of the
   meaning, it is a "Partial
   Translation".

Read the input carefully. Write
   down a brief thought process
   first, with the 4 notes in
   mind. Then identify the
   translation category of the
   idiom.
Respond in a JSON format: the
   keys must be "Category", and
   the values must be the
   corresponding category of the
   idiom translation.
```