Measuring the Literalness of English to Chinese Translations by GPT4

Xiao Wang

xiao.wang@student.uni-tuebingen.de

Abstract

This paper examines the issue of literalness in GPT4's English-to-Mandarin translations, using data from the WMT2023 Metrics Task. Despite the dataset's small size, the results reveal distinct differences in the degree of literalness across various sources. Reference translations tend to be less literal compared to machine-generated outputs, with Google producing the most literal translations, followed by GPT-4, and DeepL being the least literal. By applying the "Translation Correspondence Rate" (TCR) metric, the study quantifies these variations and offers insights into improving GPT4' s translation accuracy and naturalness.

1 Introduction

This paper presents an investigation into the systematic translation errors made by GPT4 when translating English to Mandarin Chinese using the data from the WMT2023 Metrics Task. Despite GPT4' s powerful performance in various language processing tasks, its translations in this context reveal several significant shortcomings. Drawing from the author's expertise as a native Chinese speaker with advanced English proficiency, the analysis highlights key issues, including the suboptimal quality of the reference translations and the tendency for GPT4's translations to sound unnatural and out of context. This is largely attributed to an overly literal approach, where the nuances and fluidity of the target language are sacrificed in favor of a more rigid, word-for-word rendering.

To move beyond anecdotal observation and provide a more objective evaluation, a statistical method known as "Translation Correspondence Rate" (TCR) (Imamura et al., 2003) was employed. This metric measures the literalness of translations by analyzing the alignment of words between the source and target languages. The method involved building an English-to-Chinese dictionary using the *Fastalign* automatic word alignment tool, which then enabled the calculation of TCR by counting word correspondences between the source and target sentences. The TCR is computed by dividing twice the number of aligned word pairs by the total number of words found in both the source and target sentences within the dictionary. A higher TCR indicates a more literal translation. This analysis provides a quantitative way to assess how literally GPT4 translates source text at a word level, offering insights into the underlying issues with its translation strategy.

By conducting this study, the author aims to offer insights for improving GPT4 translations. The findings underscore the need for more sophisticated handling of linguistic context, idiomatic expressions, and cultural subtleties in machine-generated translations. By addressing these gaps, future iterations of GPT4 and similar models could produce translations that are more fluid, natural, and contextually appropriate. The detailed results of the paper are available in this repository.

2 Data

This paper uses two sources of data: translation data to be analyzed and a corpus used to train word alignments for the dictionary.

The translation data comes from the generaltest2023 dataset provided by the WMT2023 Metrics Task[1]. It includes 2,074 English source sentences, along with their Chinese reference translations and machine-generated translations from GPT4, DeepL, and Google

[1]See the link: https://wmt-metrics-task.github.io/

	<.5	<.6	<.7	<.8	<.9
GPT4	1	19	73	312	1242
DeepL	0	18	91	343	1261
Google	2	12	67	266	1053

Table 1: Counts of sentences where GPT4, DeepL, and Google translations received COMET scores below 0.5, 0.6, 0.7, 0.8, and 0.9 compared to reference translations.

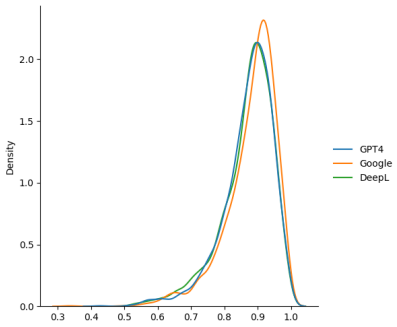


Figure 1: COMET scores comparing the translations generated by GPT4, DeepL, and Google against the reference translations.

Translate. Notably, six duplicate sentences appear after punctuation is removed, but these duplicates were intentionally retained to explore potential translation variations. While all three machine systems produced identical translations for the duplicates, the reference translation handled one duplicate sentence differently from its original.

To quickly identify poorly translated sentences, COMET scores (Crosslingual Optimized Metric for Evaluation of Translation) from the dataset were used instead of BLEU scores, as COMET scores are believed to align better with human judgment. COMET scores reveal that the dataset is unbalanced in translation quality (relative to the reference) across all three systems, with more than 83% of translations receiving scores above 0.8. For instance as shown in table 1 and Figure 1, GPT4 produces high-quality translations, with only 73 sentences receiving COMET scores between 0.42 and 0.7, of which 19 scored below 0.6. In contrast, 2,001 sentences have scores above 0.7, including 832 with scores above 0.9. DeepL and Google show similar patterns except Google translations align even more closely with reference translations, as it produced fewer sentences scoring below 0.7 and more scoring above 0.8 compared to GPT4 and DeepL.

The English-to-Chinese dictionary (word alignments) was trained using the English-Chinese parallel corpus from the WMT 2023 Machine Translation Task. Due to the author's computer limitations, only data from OPUS (Open Parallel Corpora) was utilized. The training dataset is 5.09 GB and contains 17,451,546 sentence pairs. The alignment tool *Fastalign*[2] was used to extract word alignments between English and Chinese. Before alignment, English words were converted to lowercase and English sentences were tokenized using *word_ tokenizer* from the Python package *nltk*, while Chinese translations were segmented using the package called *jieba* [3].

To improve alignment accuracy, training was performed in both directions—English to Chinese and Chinese to English—running five iterations each. The intersections of the resulting alignments were used, yielding around two million non-unique word pairs. To refine the dictionary, the top 500,000 most frequent word alignments (occurring at least seven times) were selected. For each English word, the top 10 most frequent aligned Chinese words were included, as multiple Chinese words can align with a single English word. Ideally, only the most frequent Chinese equivalent of each English word should be retained. However, due to the dictionary's quality, additional matches were included to ensure a sufficient number of linked words for calculating correspondence rates (see section 3). This, in turn, may complicate the distinction of literalness between translations of the same word, as multiple options can appear in the dictionary definition of the source word. Ultimately, this process produced an English-to-Chinese dictionary containing 158,139 entries.

It should be noted that despite the filtering applied, the dictionary created in this manner is far from perfect. Punctuation marks were not removed before alignment, so some frequent punctuation marks still appear in the

[2]https://github.com/clab/fast_align

[3]The "best" Chinese segmentation tool:https://github.com/fxsjy/jieba

final dictionary. Additionally, a small portion of misaligned words remain, primarily due to structural differences between English and Chinese. This is especially noticeable with frequent function words unique to each language, such as "the" and "to" in English, and 的 and 了 in Chinese. One of the most prominent issues is the misalignment of the English determiner "the" which has no direct equivalent in Chinese, leading to its frequent alignment with the frequent Chinese particle 的 [4]. This issue can result in different TCR scores for translations with a small difference of a function word, such as [换, 好, 的, 价格] and [换, 好, 价格] (both means "change for a good price", "for" is aligned with 的 in the dictionary). However, since such cases are infrequent and the differences are minor, they do not significantly affect the comparison of TCR scores between two machine translations over a large dataset. Nonetheless, it does affect the precise measurement of literalness for a translation when TCR is calculated using this dictionary.

3 The Analysis

The analysis of GPT4's translation quality will be presented not only by assessing its own output but also in comparison to translations produced by DeepL and Google Translate. The data was sorted based on the COMET scores of GPT4's translations relative to the reference translations. Note that sentences with low GPT4 COMET scores are not necessarily the same ones with low COMET scores for DeepL or Google translations. The primary focus of the review is on the translations of sentences whose GPT4 translations received COMET scores below 0.75. During the manual evaluation, some issues with the reference translations were also identified. A key observation is that, in comparison to reference translations, the machine-generated translations, including those by GPT4, tend to be more literal. Specifically, based on the author's evaluation, Google is the most literal, followed by GPT4, while DeepL is the least literal among them. To validate this, a quantitative measure called the "translation correspondence rate" (TCR) was used to assess the degree of literalness in the translations. The results indicate machine translations are notably more literal than the reference translations. However, as GPT4 COMET scores increase, this difference becomes less pronounced. These findings will be discussed in detail in the following subsections.

[4] 的 is commonly used at the end of adjectives, as in 漂亮的 ("pretty"), in possessive pronouns or noun phrases like 我们家的 ("our family's"), or as a sentence-final particle in 是... 的 constructions.

3.1 The Reference Translations

Upon reviewing the machine translations of sentences with low GPT4 COMET scores, it was observed that some of the low scores can be partially attributed to the poor quality of the reference translations (while the machine translations are good). The issues with the reference translations appear to arise from a lack of contextual understanding, likely due to insufficient domain knowledge or, in some instances, inadequate English proficiency.

(1) **Source** : #flutter is a pleasure to work with, and they have #mobx for easy store Management.
Reference : #flutter 是一个绝佳合作伙伴，他们拥有可以简化商店管理 的 #mobx。

(2) **Source** : Intersex is even recognized by TERFs and the like because it's generally more physical and, as such, visible.
Reference : 双性人竟然被 TERF 之类的机构所认可，因为它通常更具物理性，因此也更明显。

(3) **Source** : Casper the #rat is an absolute unit these days.
Reference : Casper the #rat 现在是一个极好的单位。

For example, as shown in the underlined phrases in example (1), the word "store" in "store management" clearly refers to data storage, given that the context involves "Flutter", an open-source UI (user interface) framework for building platform applications. However, the reference translation fails to consider this context, translating it as "shop management" (商店管理), which refers to managing physical stores. A similar issue occurs in example (2), where the sentence mentions "intersex people are more 'physical'", intending to describe the "physical" as relating to the body rather than

the mind. Instead, the reference translation incorrectly interprets it in terms of "physics" (the underlined part 物理性). Another example, example (3), highlights a lack of English proficiency in the reference translation. The slang "a unit", used to describe someone as "very large and impressive" (Cambridge Dictionary), was mistranslated using its literal meaning 单位, which refers to "a single thing or a separate part of something larger" (Cambridge Dictionary). These are just a few examples, and similar mistakes can be found throughout the dataset.

The poor quality of the reference translations presents a significant issue that needs to be addressed. Not only does it undermine the reliability of evaluation metrics like COMET scores, but it can also lead machine translation systems to produce flawed outputs, as they are often designed to optimize translations to match the reference texts as closely as possible.

3.2 The GPT4 Translations

If the issue of "lack of attention to context" is observed in some reference translations, it is prevalent in machine-generated translations from GPT4. Context, which can include syntactic cues, document-level information, or general world knowledge, is often overlooked in GPT4. This tendency can be explained by the fact that GPT4 is a probabilistic model, which favors alignments with the highest likelihood based on its training data. Consequently, it tends to select translations that are statistically frequent rather than those most appropriate given the specific context.

As a result, the most frequent Chinese mappings for English words and phrases are chosen in the translations. This completely ignores the fact that the same word or phrase can fall into different semantic categories and refer to distinct concepts depending on its framing (context). In some instances, this oversight may only make the translation sound "weird" or "off", such as translating the "wall" of a room (墙壁) as the "wall" of a garden (围墙). However, in more serious cases, it can lead to outright mistranslation. For example, in example (2), the word "like" is incorrectly translated as "be fond of or enjoy something" (喜欢 in GPT4's translation) when it actually means "such as" (比如 in the reference translation). Another instance is found in (1), where the phrase "can't" in "can't be repeated" can either mean "it's impossible to do it again" or "it should not happen again," depending on the context. Based on the surrounding sentiment in the text, the latter meaning is more appropriate, as reflected in the reference translation 不能重蹈覆辙 ("cannot let it happen again"). However, GPT4's translation interprets it as "it's impossible to be repeated" (无法重复的) by choosing its more frequent mapping, leading to a significant misunderstanding.

(1) **Source** : *(The ratings–seeking delusional irresponsibility of it all.)* It can't be repeated.
Reference : 不能重蹈覆辙。
GPT4 : 这是无法重复的。

(2) **Source** : Like krauty moods and other worldly pop music and radiophonic workshop and all sorts.
Reference : 比如 krauty moods、其他通俗流行音乐、广播电台直播间等各种形式。
GPT4 : 喜欢krauty 的情绪和其他世界的流行音乐，以及无线电工作室和各种各样的东西。

Apart from instances of mistranslation, GPT4 tends to produce outputs that, while technically correct in meaning, sound robotic or unnatural. This is particularly noticeable when dealing with English fixed expressions, such as slang, idioms, or proverbs, and in translating English into Chinese idiomatic expressions. This indicates that GPT4 not only tends to map the most frequent Chinese equivalent to an English word without properly considering the context, but also often translates sentences word by word, neglecting the semantic nuances and non-compositionality of idiomatic expressions. The reference translations, however, are rich in Chinese four-character fixed idioms in comparison, which adds depth and cultural authenticity to the translations.

For example, in (1), the English business idiom "throw something over the fence", which refers to passing a task or problem onto someone else without resolving it or providing enough context, was translated literally by

GPT4. The word "fence" was translated as 篱笆 (which refers to physical "fence" in Chinese), whereas the reference translation avoided a literal rendering to better convey the intended meaning.

This challenge becomes even more significant when translating into Chinese, a language rich in fixed idiomatic expressions. In many cases, translating English word-for-word into Chinese instead of using established Chinese idioms fails to capture the idiomatic nature of the target language, making the Chinese version feel unauthentic or awkward. For instance, in (2) and (3), two common Chinese idioms, 一分钱一分货 (literally, "one cent for one cent' s worth of goods") and 名不副实 (literally, "the name does not match the content"), were used in reference translations in favor of more literal renderings by GPT4. Both idioms are frequently used in daily Chinese conversation, and their absence in favor of more literal translations can make outputs sound less fluent.

(1) **Source** : Throwing everything you've got over the fence in response to GPT is not it.
Reference : 把你所有的一切都扔掉去回复 GPT 似乎适得其反。
GPT4 : 对 GPT 的回应，不是把你所有的东西都扔过篱笆。

(2) **Source** : Horrible product, I guess I get what I paid for...
Reference : 垃圾产品，一分钱一分货...
GPT4 : 糟糕的产品，我想我得到的就是我所付出的...

(3) **Source** : Horrible product, misrepresented
Reference : 垃圾产品，名不副实。
GPT4 : 糟糕的产品，被误导。

The above analysis is based on the author's human observations. To determine whether literal translation is a consistent problem in GPT4 translations, a statistical method is necessary for validation. To quantify the literalness of the translations, the metric called "Translation Correspondence Rate" (TCR) (Imamura et al., 2003) was applied . This approach allows for a more objective measurement of the degree to which translations follow a word-for-word pattern, and it will be elaborated in the next section.

3.3 The Literalness of the Translations

3.3.1 Translation Correspondence Rate (TCR)

To determine whether GPT4's tendency to translate words literally without considering context is a systematic issue, this paper uses the "Translation Correspondence Rate" (TCR) to measure the literalness of sentence Chinese translations. TCR was first introduced by Imamura et al. in 2003 to help select "appropriate bilingual sentences for machine translation" (Imamura et al., 2003). Imamura et al. argued that parallel corpora with context-dependent translations often introduce more incorrect transfer rules than those with literal translations in machine translation systems. Their findings showed that systems trained with bilingual corpora containing higher TCR scores produced moderately better outputs, demonstrating that TCR is an effective metric for assessing literalness, making it an appropriate tool for this study.

Before calculating the Translation Correspondence Rate (TCR), an English-to-Chinese dictionary was prepared by automatically extracting word alignments using the word alignment tool *Fastalign*, as described in section 2 on data collection and preparation. Preprocessing steps included removing both English and Chinese punctuation marks besides converting all English words to lowercase. This was necessary because the dictionary contained punctuation alignments, but only word alignments were relevant for TCR calculation. The same English tokenization tool and Chinese segmentation tool used during the automatic alignment process were applied once again to tokenize and segment the source and translated sentences before calculating the TCR.

TCR involves three key components:

- *Ts* represents the number of source words (English words) from the source sentence found in the dictionary.
- *Tt* denotes the number of target words (Chinese words) from the translation sentence found in the definition parts of the dictionary.

TCR Calculation Examples		
	TCR Values	Examples
Ref.	$\frac{2*1}{10+3} = 0.154$	垃圾, 产品, 一分钱, 一分货
Source	—	*horrible, product, i, guess, i, get, what, i, paid, for*
GPT4	$\frac{2*8}{10+13} = 0.696$	糟糕, 的, 产品, 我, 想, 我, 得到, 的, 就是, 我, 所, 付出, 的
GPT4	$\frac{2*4}{4+4} = 1.0$	更换, 的, 好, 价格
Source	—	*good, price, for, replacement*
DeepL	$\frac{2*1}{4+2} = 0.333$	更换, 价格合理
Ref	$\frac{2*0}{0+0} = 0.0$	person4, unintelligible
Source	—	*person4, unintelligible*
Google	$\frac{2*0}{0+1} = 0.0$	人物, 4unintelligible

Figure 2: Three source sentences and their two different translations as well as the TCR scores of the translations. Words that are underlined are found in the dictionary. Linked words are connected with lines.

- L is the count of linked word pairs where the English source word and its corresponding Chinese target word are matched based on the dictionary definitions.

The correspondence rate for a parallel sentence pair is thus calculated by multiplying the number of linked words by two, and then normalizing this value by the total number of source and target words found in the dictionary. This is expressed in the following equation (Imamura et al., 2003):

$$TCR = \frac{2L}{Ts + Tt}$$

The design of TCR effectively captures the proportion of directly (literally) translated words relative to all words requiring translation. By including both source and target words found in the dictionary in the denominator, it becomes bidirectional and accounts for differences from both sides. Additionally, it remains unaffected by reordering in translation, as the target word corresponding to a source word, regardless of its position in the target sentence, still counts as one linked word pair.

TCR values range from 0 to 1, with higher values indicating a greater correspondence rate, or a higher degree of literalness. A value of 0 indicates either no linked words (L=0) or, more extremely, when neither source nor target words are found in the dictionary (Ts=Tt=0). Conversely, a value of 1 occurs when the number of source words in the dictionary matches the number of target words, with each word forming a one-to-one correspondence in linked word pairs, represented by the formula $\frac{2*n}{n+n}$.

One notable issue is the retention of English words in the translations. This primarily occurs with common English nouns, such as place names, institution names, and acronyms. While this practice is not ideal, it is somewhat acceptable, especially for terms so frequently used in their original form that they are even recognized by the word aligner and survived in the final dictionary. However, English words that are rarely kept in translations (and thus absent from the result of the aligner) which are often those that should not be retained in the translation were not counted as linked words. In fact, there were two cases where entire source sentences were left untranslated, and their TCR scores were 0 as a result. Some examples of translations with their TCR scores are given in Figure 2.

3.3.2 The Literalness Results

The study compares and analyzes their Translation Correspondence Rate (TCR) scores across four sources: reference translations,

GPT4, DeepL, and Google Translate. The main focus was on the 174 sentences where GPT4 translations received COMET scores below 0.75, as lower scores are more indicative of potential translation issues.

The comparison and difference analysis centers on the various translations of the same sentences. Since the primary aim of this study is to assess GPT4's translation quality, its translations were used as the baseline for comparison against other systems. To measure their differences, the TCR scores of reference translations, DeepL, and Google translations were subtracted from GPT4's TCR scores for the same sentences. The resulting values range from -1 to 1, with the zero, negative, or positive values indicating whether the literalness of GPT4's translation is similar to, higher than, or lower than that of the other systems for the same sentences. Additionally, this comparison was extended to sentences where GPT4's COMET scores fell between 0.7–0.75, 0.75 – 0.8, and 0.8 – 1.0 to explore how literalness trends shift with better-performing GPT4 translations.

By analyzing the results of the 73 sentences where GPT4's COMET scores are below 0.7 (see plot (a) in Figure 3 and statistics in table 2), it is evident that the reference translations exhibit more negative values compared to those of DeepL and Google, as reflected by the negative median of the reference-GPT4 translations difference distribution and the ratios of negative TCR values in the distribution. In contrast, the medians of DeepL and Google translations are positive and closer to zero. This suggests that the reference translations for these 73 sentences tend to have smaller TCR values (less literal) than those generated by GPT4, while DeepL and Google translations show a similar and even slightly higher level of literalness to GPT4 with their positive medians and higher percentage of positive TCR difference values. Furthermore, the larger standard deviation of the reference translations indicates significantly greater variation in literalness between the reference and GPT4 translations, compared to DeepL and Google. In contrast, DeepL' s and Google's TCR differences with GPT4 cluster closely around the medians (nearly 0), as shown by their sharp, high distribution peaks and their small ranges of values for the middle 50%, further highlighting the similarity in literalness between DeepL and GPT4, as well as between Google and GPT4 for these lower COMET-scoring sentences. This narrow gap in literalness between the translations of DeepL and GPT4, as well as between Google and GPT4, combined with the greater variability observed between the reference translations and GPT4, leads to the conclusion that reference translations are generally less literal and more varying than the machine-generated translations for these 73 sentences that are poorly translated by GPT4. However, no salient difference between the literalness of DeepL and Google's translations are observed on the same sentences except that DeepL has an even smaller variance.

GPT4 COMET	Total	Ref.	DeepL	Google
<0.7	73	.56	.37	.42
0.7-0.75	101	.60	.53	.44
0.75-0.8	138	.53	.57	.42
>0.8	1762	.53	.51	.41

Table 2: Ratios of sentences where reference, DeepL and Google translations have a **negative** TCR difference compared to GPT4 translations. Sentences are grouped by GPT4 translations' COMET scores: below 0.7, 0.7-0.75, 0.75-0.8, and above 0.8. "Total" represents the total sentence count in each group.

The observed tendency of reference translations being less literal and more diverse than machine translations persists for sentences with GPT4 COMET scores above 0.7. From (b), (c) and (d) in Figure 3 and the ratios in table 2, it is clear that reference translations still show more negative TCR differences compared to GPT4 than positive or zero differences (ratio of negative TCR > 0.5). Additionally, their TCR differences continue to display greater variance than those of DeepL and Google, reinforcing the idea that reference translations exhibit the least literalness and the greatest variety overall. However, with higher COMET scores (over 0.7), some notable changes occur:

1. The gap between reference and machine-generated translations narrows, as the distribution of reference translations shifts right, aligning its median more closely

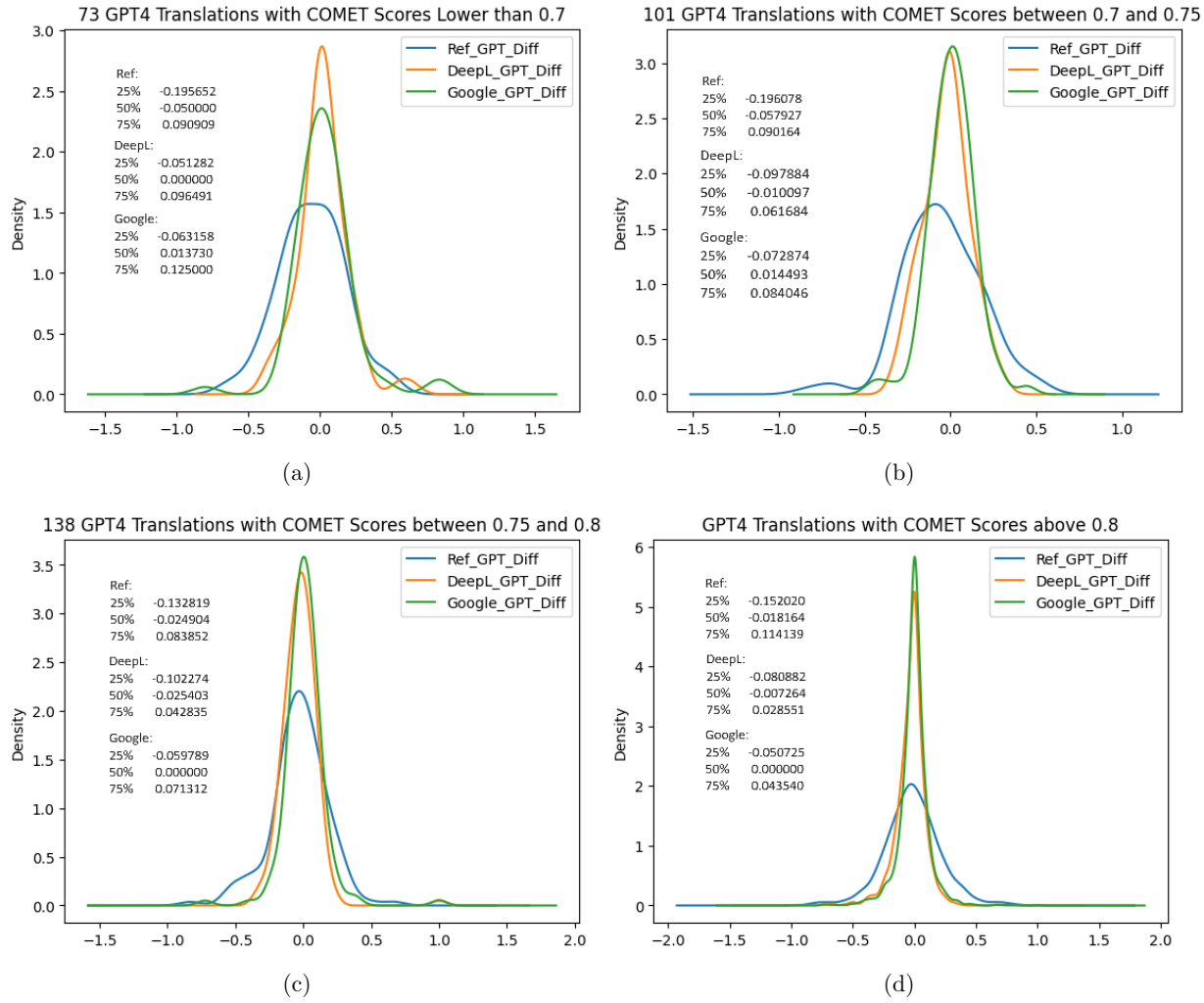


Figure 3: TCR score differences of the reference translations, DeepL translations, and Google translations compared to GPT4 for sentences whose GPT4 translations' COMET scores below 0.7 (a), between 0.7-0.75 (b), 0.75-0.8 (c) and above 0.8 (d). 50% refers to the median value.

with 0.0 (although still negative) like the others.

2. The differences between DeepL and GPT4, as well as between Google and GPT4, decrease, reflected by their decreased variances (for Google, also by its median moving further closer to zero).

3. A divergence between DeepL and Google emerges, with DeepL-GPT4's difference distribution shifting left from that of Google (results in a negative median) and increasing its ratio of negative TCR values to above 50%, indicating that DeepL's translations become less literal compared to both GPT4 and Google.

In conclusion, based on the TCR differences from GPT4 translations, reference translations consistently differ from machine-generated translations across the dataset, showing a lower degree of literalness and greater variety. DeepL and Google, on the other hand, closely match GPT4 in literalness, with near-zero medians and small variances. As GPT4' s COMET scores rise, this trend holds, except for DeepL, which shows a decreasing literalness compared to GPT4 and Google. The results indicate an overall literalness ranking: reference, DeepL, GPT4, and Google.

3.3.3 Discussion

Although the overall results align with the manual observations of the author that GPT4 translations are generally more literal than reference translations and DeepL, and Google translates most literally, the differences are

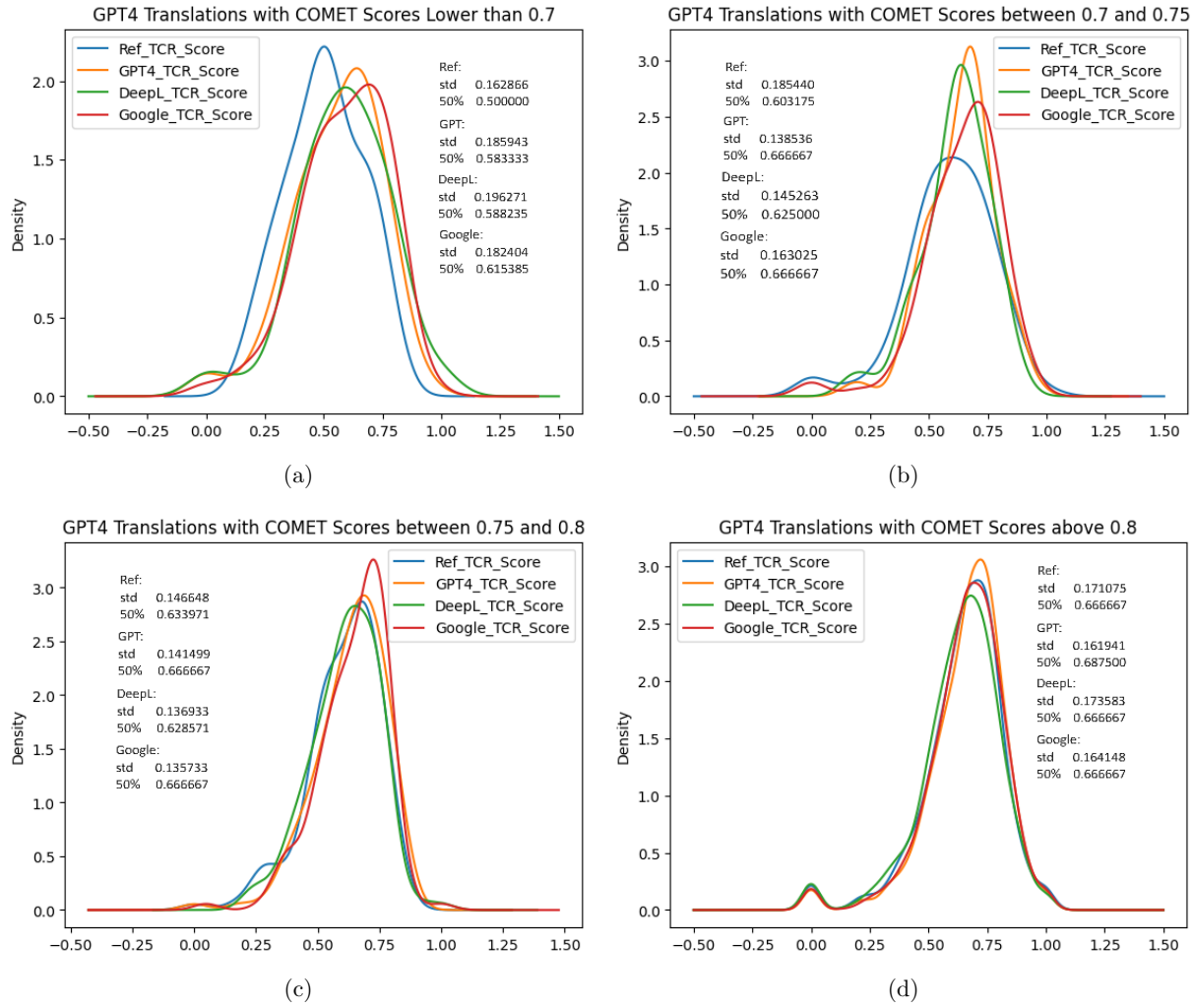


Figure 4: TCR scores of the reference translations, GPT4 translations, DeepL translations, and Google translations across sentences whose GPT4 translations' COMET scores below 0.7 (a), between 0.7-0.75 (b), 0.75-0.8 (c) and above 0.8 (d). *std* refers to the standard deviation, and 50% refers to the median value.

small and some unexpected patterns show.

The minimal differences in literalness are evident in the TCR comparisons between reference and GPT4, DeepL and GPT4, and Google and GPT4 translations with their large areas of distribution overlapping as shown in Figure 3, which can be further confirmed with their overall TCR score distributions. Examining the TCR scores of all sources across different GPT4 COMET score groups in Figure 4, it is also observed:

- Although reference translations consistently show TCR values distributed slightly to the left of the others across all groups, indicating the lowest level of literalness, the overall differences in TCR scores between the reference translations, GPT4, DeepL, and Google are small, as evidenced by their nearly overlapping distributions. And these differences diminish even more as the COMET scores of GPT4 translations increase, aligning with the findings in TCR difference Figure 3.

What's more, the distributions of TCR values in Figure 4 further demonstrate that GPT4's COMET scores and the TCR scores are correlated:

- More than half of the translations from all sources show literalness above 0.5 (with TCR medians higher than 0.5), except for the reference translations in the COMET score group below 0.7, where the median is exactly 0.5.
- The overall literalness rises with higher GPT4 COMET scores, as indicated by

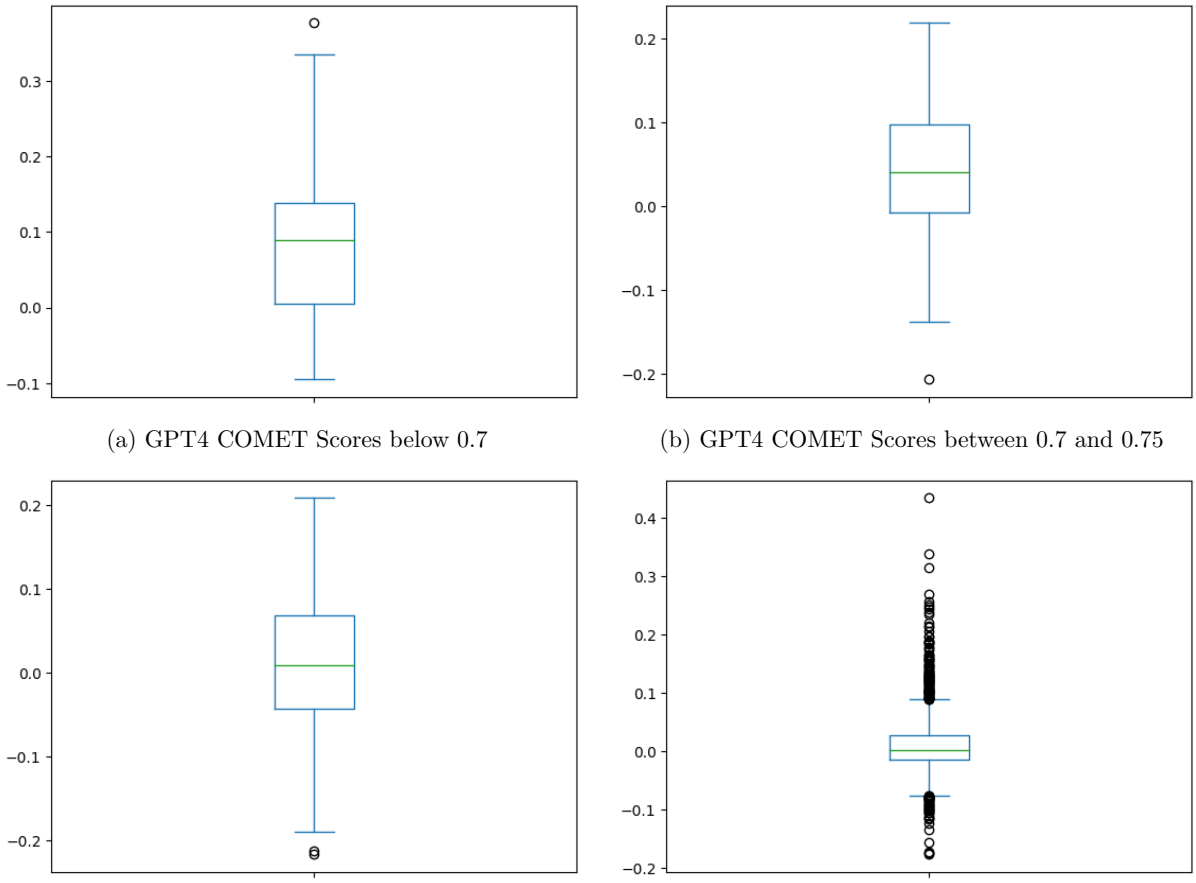
the rightward shift in distributions and increasing medians.

The unexpectation is that for the 73 sentences with lower GPT4 COMET scores (below 0.7), GPT4 translations did not show higher literalness compared to DeepL, whereas translations with higher COMET scores (between 0.7 and 0.8) did. One possible explanation could be that DeepL translations of these sentences generally have higher COMET scores than GPT4 translations. Since it is observed that higher COMET scores correlate with increased literalness as shown in Figure 4, the higher COMET scores for DeepL translations could account for their higher literalness compared to GPT4. Figure 5, a box plot of the differences between DeepL's and GPT4's COMET scores for the same sentences, shows that this gap widens for sentences where GPT4's COMET scores fall below 0.7, with a range of (-0.09, 3.4) and a median of 0.89 compared to a narrower range (-1.3, 2.2) and a smaller median 0.40 when GPT4's COMET scores are between 0.7 and 0.75. As the COMET score differences between DeepL and GPT4 reduce for sentences where GPT4's COMET scores range between 0.7 and 0.75, DeepL's tendency to translate less literally than GPT4 becomes more apparent. From the plot (c) and (d) of Figure 5, we can also see that as GPT4 COMET scores increase, the differences between the COMET scores of DeepL and GPT4 for the same sentences further decrease. However, this does not rule out the possibility that GPT4's low COMET scores (below 0.7, indicating poor translations) may be influenced by factors beyond literalness, or that the small dataset of 73 sentences is insufficient to capture these differences. More data is needed to investigate further.

Finally, it is important to acknowledge the limitations of the study. The small size of the dataset, the poorly translated sentences from GPT4, is a key issue; results derived from a larger and more diverse dataset would be more revealing and reliable. Additionally, data preprocessing could be enhanced to improve word alignments by better handling stop words and removing punctuation. Employing a more effective Chinese segmentation tool, if available, or manually correcting misaligned segments would also be beneficial. Improved word alignments can reduce misaligned word correspondences, subsequently improving the quality of the dictionary. Applying harsher filtering could further improve the dictionary's quality. A richer and more accurate dictionary would more accurately reflect the true literalness of the translations.

References

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003. Automatic construction of machine translation knowledge using translation literalness. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.



(a) GPT4 COMET Scores below 0.7

(b) GPT4 COMET Scores between 0.7 and 0.75

(c) GPT4 COMET Scores between 0.75 and 0.8

(d) GPT4 COMET Scores above 0.8

Figure 5: COMET score differences between DeepL and GPT4 translations across sentences whose GPT4 translations' COMET scores below 0.7 (a), between 0.7-0.75 (b), between 0.75-0.8 and above 0.8. The values = (DeepL COMET scores - GPT4 COMET scores) of the same sentences.