International Journal of Language and Literary Studies
Volume 7, Issue 2, 2025
Homepage : http://ijlls.org/index.php/ijlls

From Text to Understanding the Inner Text: LLMs and Translation Accuracy and Fluency

Mahdi R. Aben Ahmed
Associate Professor of TESOL and Applied Linguistics, Jubail English Language and Preparatory Year Institute, Royal Commission for Jubail and Yanbu, Jubail, Saudi Arabia
Mha1121@hotmail.com

DOI: http://doi.org/ 10.36892/ijlls.v7i2.2072

APA Citation: Aben Ahmed, M. (2025). From Text to Understanding the Inner Text: LLMs and Translation Accuracy and Fluency. *International Journal of Language and Literary Studies. 7*(2).139-156. http://doi.org/10.36892/ijlls.v7i2.2072

Received:
15/01/2025

Accepted:
16/02/2025

Keywords:
Accuracy, Error analysis, Fluency, LLMs, Machine Translation, Naturalness.

Abstract

Evaluating translation quality is crucial to ensuring that Large Language Models (LLMs) meet the ambitious standards required for effective communication across languages. The key aspect of translation quality is accuracy; accuracy measures how well the translation reflects the meaning of the original text. It also measures the fluency based upon the naturalness and readability of the translation in the target language, both features play a crucial role in defining what smooth translation should appear to the prospective user(s). The present study, therefore, aims to measure these aspects of LLM-generated translation based on a corpus of LLM-translated texts. As this study is evaluative, it proposes a rigorous method to evaluate and improve the accuracy, fluency, and naturalness of LLM-generated translation. The study, therefore, analyses AI-generated translation texts based on these criteria. The secondary data set was collected from users of AI-based translation to provide further insights into the validity of LLM-based translation texts. By combining both real-time translated texts generated by AI and users' perspectives, this study arrived at results on the status and validity of LLM-based translation. The majority of the participants concurred that the translations retained the meaning of the original text, even the lower scores for processing idiomatic expressions and figurative language in LLMs still reflected a high level of semantic preservation, The high scores for grammatical correctness and sentence flow show that LLMs are perceived to be very good at generating translations that are grammatically correct and readable. Based on the findings, the study offers practical implications for future enhancement in AI-generated translation.

1. INTRODUCTION

Adequate translation should reflect three main constructs: it should be accurate, clear, and natural. According to Anari (2004), while adequate translation suggests that translated texts maintain the intended meaning of the original texts without causing any distortion, clear translation means that translated texts should be clear. Thus, natural translation indicates that translated texts should be natural-like in the target language. As stressed by Alves et al. (2022), these three qualities of translation are necessary for the effective communication of ideas and meanings in specific fields such as healthcare, international business, and trade. Misinterpretations in such fields are liable to cause serious damage, including life risks, especially in sensitive and critical situations. Therefore, enhancing precision and smoothness is the main objective of integrating the latest technologies in the field of translation.

Large language models, also known as LLMs, represent a feasible solution in translation technology in this context. In this regard, Lyu et al. (2023) point out the main reasons behind the power of these models. These models can address certain crucial aspects of translation quality. In addition, they can leverage deep learning and absorb vast amounts of data to learn the intricacies of language(s) as opposed to traditional translation models, which are heavily reliant on statistical approaches and restricted rules. Qian et al. (2021) add that modern translation models, such as ChatGPT-3 and BERT, are well-trained in diverse datasets, including a variety of linguistic contexts, cultural nuances, and idiomatic expressions, which enable them to translate texts that may not be totally accurate, but fluent (e.g., capturing the subtleties) and natural (e.g., reflecting the flow of the target language).

Despite the advancements made by LLMs, they are still not perfect in generating accurate translations. This is mainly because of failures to manage ambiguities and tricky context-specific translations. According to Iyer et al. (2023), different languages contain words and phrases whose meanings vary depending on the context where they are used. For example, in English, the noun 'bank' might be used to mean either an institution or the side of a river. Therefore, for LLMs to produce accurately translated texts, they need to be trained in understanding these linguistic uncertainties. However, this may turn out to be particularly challenging for these models to understand the contexts in which these words are used, especially when dealing with disorganized sentences. Hence, overcoming these challenging issues resulting from using LLMs in translation necessitates model architecture, constant refinement of training data and diversity of assessment techniques.

One interesting method to address these challenges is to incorporate more diverse and context-rich datasets that enable these models to handle linguistic and contextual ambiguities better. Implementing mechanisms for detecting and exploring potential inaccuracies in using LLMs in translation is also necessary for helping users have greater confidence in LLM-generated translations. Furthermore, advancements in real-time translation technology powered by LLMs can potentially change multilingual communication in everyday interactions, education, and international business (Jermakowicz, 2023). As LLMs are in the process of development, their ability to offer more accurate, dependable, and high-quality translations is likely to increase which can bridge language barriers and facilitate global understanding.

To conclude, while LLMs have significantly enhanced the fluency and naturalness of machine-generated translations, challenges related to accuracy, especially in handling contextual ambiguities, remain a chief concern. Addressing these limitations requires continuous refinement of model architectures, expansion of training datasets, and implementation of robust assessment techniques. By integrating diverse linguistic contexts and improving error-detection mechanisms, LLMs can move closer to achieving high-quality, precise, and contextually appropriate translations. As these models continue to evolve, they hold great promise in breaking language barriers and fostering more effective global communication and cooperation.

2. LITERATURE REVIEW

As mentioned earlier, accuracy and fluency are two essential pillars of translation quality (Graham et al., 2019). In contrast, accuracy ensures that the translated text faithfully reflects the original content without missing words, unnecessary insertions, or distortions of the intended meanings. Fluency represents how translation in the target language becomes natural and grammatically acceptable. Previous researchers agree that achieving a balance in these two important aspects of translation is a challenge because the enhancement of one aspect may result in inevitable adjustments in the other aspect (Toral et al., 2018). For example, an LLM may be able to translate a sentence that is grammatically acceptable and fluent. However, this sentence may not necessarily convey the source text's meaning in an accurate manner, thus further stressing the need for rigorous assessment metrics that consider both aspects simultaneously.

Yet, numerous studies have shown that the advent of LLMs has transformed the field of natural language processing; LLMs have demonstrated remarkable competencies in understanding and generating human-like text (Mahdy, Samad & Mahdi, 2020;Yadav, 2024). Traditional machine translation systems, such as rule-based and statistical machine translation, relied heavily on predefined linguistic rules and large parallel corpora, which often struggled with idiomatic expressions and complex sentence structures (Naveen & Trojovský, 2024; Al-Wasy & Mohammed, 2024).). One of the most significant advantages of LLMs in translation is their ability to understand and maintain context over long passages (Muñoz Andrés, 2024; Richards & Martinez, 2024). Unlike traditional systems that translate sentences in isolation, LLMs can consider the broader context, leading to more coherent and contextually appropriate translations (Nordin & Schmidt, 2024). On the other hand, despite of their impressive capabilities, LLMs' can inherit biases present in their training data, leading to biased or unfair translations. Also, despite their advanced level of accuracy, LLMs may still struggle with highly specialized terminology or nuanced domain-specific language (Qamar & Raza, 2024). For these issues, such AI models must be subjected to frequent evaluation.

With machine translation tools becoming an inevitable part of the translation milieu, various methods have been developed to evaluate translation quality, ranging from automated metrics to human evaluations. Papineni et al. (2002) declared that automated metrics, such as TER (Translation Edit Rate), METEOR (Metric for Evaluation of Translation with Explicit Ordering), and BLEU (Bilingual Evaluation Understudy) are commonly used for the efficiency of AI-generated translation. Yet, the primary focus of these metrics is on the overlapping between the translation and reference texts, thus failing to fully capture the nuances of accuracy and fluency (Mathur et al., 2020). As stated by Bentivogli et al. (2016), human assessment can provide qualitative insights into the grammaticality, quality, and readability of translation. However, it is more resource-intensive, thus making humans a valuable complement to automated assessments. Moreover, human evaluation is particularly crucial in assessing translations for low-resource languages or specialized domains (Lankford, 2024), where AI models may struggle due to limited training data or domain-specific nuances. For instance, in legal translation, where precision and accuracy are paramount, human evaluators can ensure that the translated text is strictly observed by following some required standards and conveying the intended meaning without ambiguity.

Since the focus of this proposed research is translation quality, error analysis becomes a critical component of quality assessment. In this regard, Isabelle et al. (2017) argues that error analysis involves the identification and categorization of errors, such as mistranslations and grammatical mistakes, which enables an understanding of where LLMs are capable of excelling and where they fail to achieve such tasks. Recently, researchers have highlighted the capability of LLMs to provide high-quality translations in common languages. However, they are faced with challenges in complex syntactic structures, especially in low-pressure languages (Koehn, 2020). This discrepancy underscores the importance of targeted error analysis for improving model performance across diverse linguistic contexts. By identifying specific areas where models struggle, such as long-range dependencies, idiomatic expressions, or morphological richness, researchers can develop more effective strategies for fine-tuning and enhancing the robustness of LLMs in managing these linguistic challenges.

Amongst previous research, case studies have provided evidence of the role of these models in achieving significantly improved quality of text translation. For example, a study comparing translations given by GPT-3, a state-of-the-art LLM, with those generated by traditional NMT systems showed that GPT-3 provides more accurate and fluent translations across multiple languages. On this, Raunak et al. (2023) claim that LLMs are capable of managing the translation of idiomatic expressions and complex sentence structures more efficiently, thus achieving contextually appropriate and natural-like translations.

Orlando, Liao, and Kruger (2024) discuss the rise of large language models (LLMs) in enhancing traditional neural machine translation (NMT) capabilities, particularly in tasks like post-editing machine translation (PEMT) error corrections. The study also emphasizes the ethical challenges associated with the use of these advanced models.

Likewise, a survey by Kalyan (2024) offers reviews on multi-modal LLMs and the safety and trustworthiness of LLMs, respectively. Nevertheless, there is currently no comprehensive survey dedicated to the GPT-3 family of LLMs. Given the increasing prominence of GPT-3 family models such as GPT-3, Instruct GPT, Chat GPT, and GPT-4, and the growing volume of research utilizing these models, there is a pressing need for a survey that specifically concentrates on the GPT-3 family of LLMs. Hadi et al. (2023) furthermore concluded the future directions of LLM research and identified key challenges that must be addressed to enhance the reliability and utility of these models.

In addition, Denecke et al. (2024) found that these models achieved highly accurate translation of specialized terminologies and complex medical information, which are both crucial for effective and accurate healthcare communication. The authors argued that these results are concrete evidence of the potential of LLMs in performing high-quality translation, especially in sectors such as healthcare, legal writing, content development and international business that emphasize precise and fluent communication of information.

From these previous case studies, using standardized corpora and benchmarks is important for consistent evaluation of the quality of LLM-generated translation. As stated by Barrault et al. (2019), benchmarks such as WMT (Workshop on Machine Translation) can offer good datasets and establish frameworks for evaluations through model comparisons. Typically, these benchmarks are inclusive of a mixture of high-resource and low-resource language pairs, which consequently enable translators and researchers to assess the performance of LLMs in translation across diverse scenarios. Standardized benchmarks play a role in promoting transparency and reproduction possibility in translation quality research. They also foster innovation and collaboration within the field (Specia et al., 2018).

2.1.Research questions

Based upon the foregoing review of literature and identification of research gap in the Saudi English-Arabic language pair context, the present study aims to answer the following questions:

3. To what extent do LLMs perform high-quality translation tasks?
4. What are the types and nature of errors in LLM-generated translations of texts?

5. METHODS

5.1.Research Design

The study uses a corpus research design to assess the quality (and errors) of LLM-generated translation. This design is suitable as it allows researchers to examine genuine tasks performed by these models. Besides, the study seeks to better understand the quality and errors of LLM-generated translation from users' perspectives. Moreover, this research utilizes a quantitative research method to evaluate the accuracy, fluency, and naturalness of large language model (LLM) produced translations. The research uses a 5-point Likert scale-based questionnaire as the main evaluation tool, enabling a systematic and organized analysis of translation quality. The questionnaire is used to measure various aspects of translation performance, such as accuracy, fluency, naturalness, consistency of terminology, error management, and comparative assessment.

Twenty professional translators, bilingual language specialists, and linguists with experience in translation quality assessment were the participants of this research. The 5-point Likert scale questionnaire was filled out by each participant based on a set of translated texts for evaluation. A total of LLM-generated translations was evaluated independently by each participant using the Likert scale questionnaire. The answers were gathered and tabulated quantitatively, with mean scores on each dimension to ascertain the overall quality of the translations.

5.2.Data Collection and Treatment

In realizing the objective of this research, the researcher locates the probable flaws of LLM-generated translation. To that end, effective data analysis is required to determine the degree of improvement in the quality of LLM-based translation concerning its quality. The research also uses benchmarks such as varied linguistic characteristics, idiomatic phrases, and domain-related content to conduct an exhaustive evaluation of the LLM-based translation’s strengths and weaknesses for the language pair under consideration. This also yields information on the role of the various linguistic and contextual parameters influencing the quality of LLM-translated text.

In this research, error analysis is another vital element of data analysis. This entails the identification and classification of errors in LLM-produced translations to learn about typical issues and areas for improvement. The identified errors are grouped into categories like semantic (e.g., incorrect word meanings), syntactic (e.g., grammatical errors), and pragmatic errors (e.g., cultural errors). By analyzing these mistakes systematically, the research will gain valuable insights into the limitations and shortcomings of LLM-based translation and propose strategies on how to enhance the accuracy and fluency of such translations.

6. DATA ANALYSIS

The examination of the 20 respondents' answers to the 30-item Likert Scale questionnaire provides a number of insights into the performance of LLMs in translation accuracy, fluency, naturalness, consistency of terminology, error management, and comparative reliability. Table 1 summarizes these results.

Table 1: Accuracy and Fluency

No	Item	Disagree	Neutral	Agree	Strongly Agree	Mean	Rank
1	The translated text accurately conveys the meaning of the original text.	3	5	8	4	3.65	5
2	The translation does not omit or add crucial information.	0	3	11	6	4.15	1
3	The translation preserves the tone and intent of the original text.	0	7	9	4	3.85	4
4	The translation maintains appropriate formality and register.	0	7	8	5	3.9	3
5	The translation correctly represents idiomatic expressions or figurative language.	0	7	7	6	3.95	2

Figure 1: Accuracy and Fluency

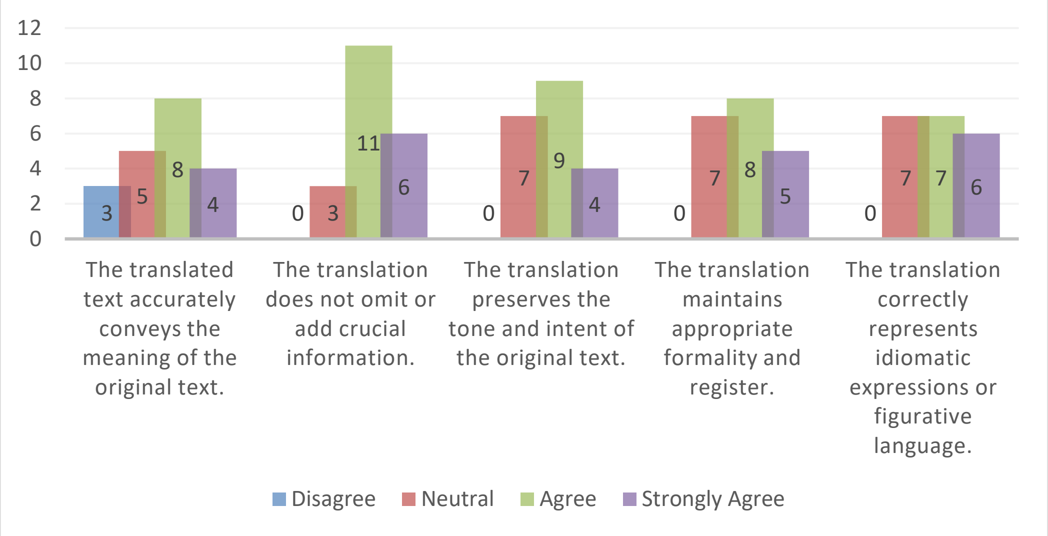

Table 1 and Figure 1 show the perceptions of the participants about the accuracy and fluency of a translated text. The statement, "The translation does not omit or add important information," had the highest ranking with a mean of 4.15, reflecting strong consensus among participants that the translation by LLM is close to the original text. This concurs with the findings. In contrast, the item ranked lowest, "The translated text accurately represents the meaning of the original text," had a mean of 3.65, reflecting dissatisfaction or uncertainty with the overall accuracy of the translation, Dahia and Belbacha (2024), highlighting that ChatGPT's translation of was in a moderate accuracy level. Also, it concurs with the observation of Mohamed et al. (2024) that translating the complete meaning of a source text frequently is difficult because of differences in languages and cultures.

The statements related to tone, intent, and formality (e.g., 'The translation preserves the tone and intent of the original text’ and 'The translation maintains appropriate formality and register') scored moderately, with means of 3.85 and 3.90, respectively. However, the significant number of neutral responses (7 for each item) suggests that respondents were less confident about these aspects, due to subtle nuances in tone or register that were not fully apprehended. This finding is consistent with = Al-Kaabi et al. (2024) which highlights the difficulty of preserving stylistic and tonal elements in translation, especially across languages with differing cultural norms. The statement addressing idiomatic expressions, scored a mean of 3.95, indicating robust performance of LLMs on this count but still leaving room for improvement, as 7 respondents remained neutral. This aligns with the research of Nazeer et al. (2024) which argues that idiomatic and figurative language often pose significant challenges in translation due to their cultural specificity.

Table 2: Fluency and Readability

No	Item	Neutral	Agree	Strongly Agree	Mean	Rank
1	The translation is free from grammatical errors.	8	8	4	3.8	4
2	The sentence structure follows natural language patterns.	6	7	7	4.05	1
3	The translation is coherent and easy to follow.	7	8	5	3.9	3
4	The choice of words is appropriate for the target language.	9	7	4	3.75	5
5	The translation is structured logically without awkward phrasing.	7	6	7	4	2

Figure 2: Fluency and Readability

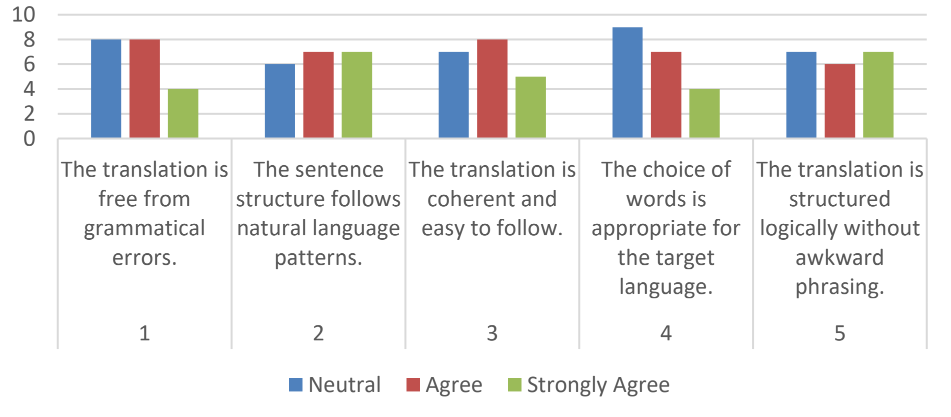

Table 2 and Figure 2 show the evaluation of the participants for the fluency and readability of a translated text from the participants' perspectives. The emphasis was placed on grammatical correctness, natural language habits, coherence, choice of words, and logical organization. The statement, "The sentence structure follows natural language habits," ranked the highest at a mean rating of 4.05, which reflected the prominent level of agreement that the translation follows the syntactic conventions of the target language.

This agrees with Mohsen (2024) who established that AI tools had the ability to grasp contextual subtleties, identify city names, and acclimate to the style of the target language. This was followed by, 'The translation is structured logically without awkward phrasing’ scored a mean of 4.00, further supporting the translation's fluency. This is consistent with the study by Davoudi Sharifabad and Rajabi Fakhrabadi (2023) which highlighted the challenges of maintaining logical flow while avoiding unnatural phrasing in translated texts.

The statement 'The translation is coherent and easy to follow' scored a mean of 3.90, indicating potential coherence issues, aligning with Olmedilla et al. (2024) which found that AI, models are still weak at this point. The lowest-ranked statement, "The choice of words is appropriate for the target language," scored 3.75, reflecting lexical uncertainty. This means that LLM still suffers from the proper word selections. These results highlight persistent difficulties in achieving coherence, grammatical accuracy, and lexical precision in translation.

Table 3: Naturalness and Human-Like Quality

No	Item	Disagree	Neutral	Agree	Strongly Agree	Mean	Rank
1	The translation reads as if written by a native speaker.	0	10	5	5	3.75	4
2	The translation avoids direct word-for-word rendering were unnecessary.	0	12	7	1	3.45	5
3	The flow of sentences is smooth and natural.	1	6	7	0	3.9	3
4	The translation maintains expressive qualities such as emotion or nuance.	0	4	9	7	4.15	2
5	The translated text does not sound robotic or overly literal.	0	0	10	10	4.5	1

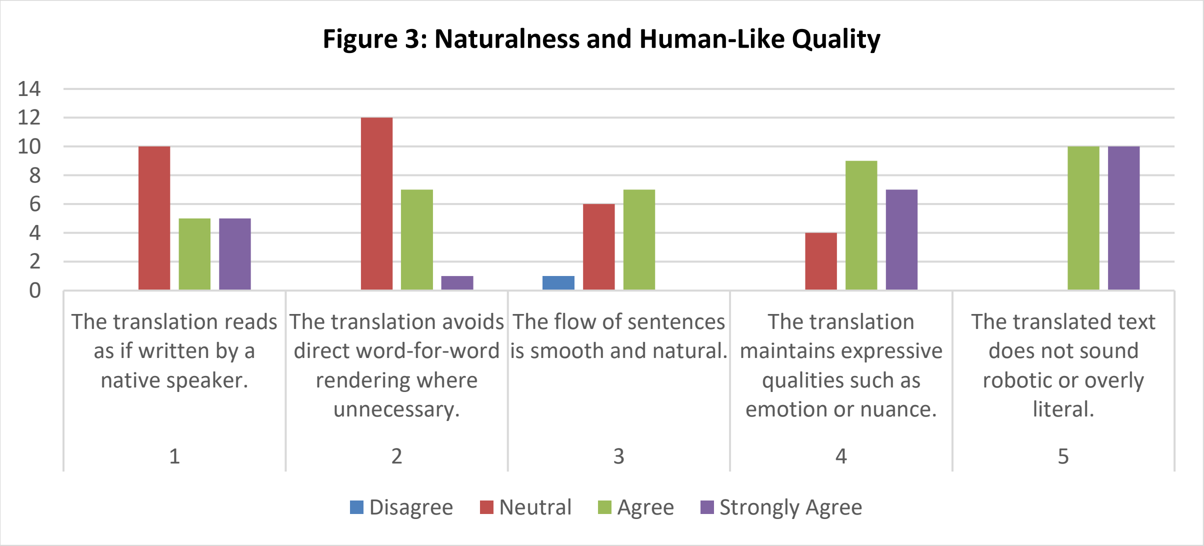

Figure 3: Naturalness and Human-Like Quality

The items in Table 3 and Figure 2 discuss the naturalness and human-like quality of a translated text, focusing on native-like fluency, avoidance of literal rendering, sentence flow, expressive qualities, and non-robotic tone. The top-ranked item, "The translated text does not sound robotic or overly literal," had a mean of 4.50, showing strong agreement that the translation has natural, human-like features, with no reported disagreements. This is in line with the findings by Jiang, et al (2024) who discovered that AI-Chat GPT generates high-quality human-like translations. The second-placed item, "The translation retains expressive features like emotion or nuance," achieved a mean of 4.15, showing excellent performance in retaining subtle aspects of the source text, although 4 respondents were neutral, indicating slight inconsistencies. Smooth sentence flow and translation sound as if it was written by a native speaker, also one such excellent feature of LLM, indicating near-native quality. The lowest-ranked entry, "The translation steers clear of literal word-for-word translation where unnecessary," received a mean of 3.45, with 12 neutral responses reflecting strong doubt regarding the translation's capacity to transcend literalness.

Table 4: Terminology and Consistency

No	Item	Neutral	Agree	Strongly Agree	Mean	Rank
1	The translation consistently uses correct terminology.	11	9	0	3.45	4
2	Key terms are translated appropriately based on the domain or context.	0	11	9	4.45	1
3	Proper names, places, and technical terms are handled correctly.	10	8	2	3.7	3
4	The translation does not introduce inconsistencies in terminology.	8	10	2	3.7	3
5	Abbreviations and specialized terms are used correctly.	0	13	7	4.35	2

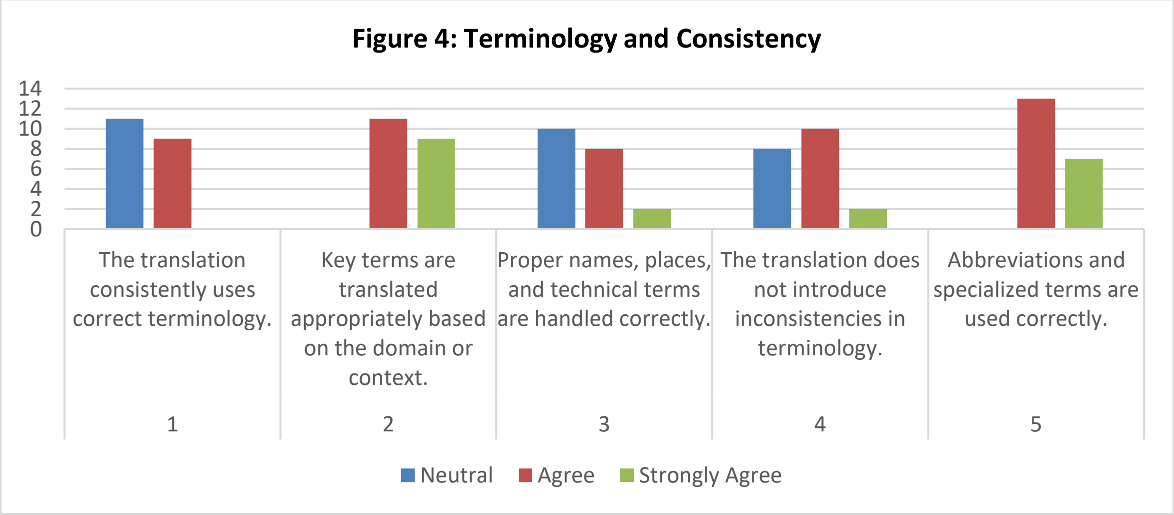
Figure 4: Terminology and Consistency

Table 4 and Figure 4 summarize the perceptions of the respondents towards the terminology and consistency in a translated text, focusing on the correct and consistent use of terminology, domain-specific appropriateness, handling of proper names and technical terms, and accuracy of abbreviations. The highest-ranked statement, "Key terms are translated appropriately based on the domain or context," achieved a mean of 4.45, indicating strong agreement that the translation effectively adapts terminology to the relevant context, amazingly with no neutral or disagreeing responses. This aligns with Falempin and Ranadireksa (2024) which highlighted AI's potential to enhance efficiency. The second-ranked statement, "Abbreviations and specialized terms are used correctly," scored a mean of 4.35, reflecting strong performance in handling technical language, though 13 neutral responses suggest some uncertainty or variability in this area.

The lowest-ranked statement, "The translation consistently uses correct terminology," scored a mean of 3.45, with 11 neutral responses, suggesting significant variability or uncertainty in the consistent application of terminology. The respondents reported earlier that LLM produces Correct terminology, but sometimes it lacks consistency.

Table 5: Error Handling and Robustness

No	Item	Neutral	Agree	Strongly Agree	Mean	Rank
1	The translation accurately interprets ambiguous or polysemous words.	5	9	6	4.05	1
2	The translation does not introduce misleading or incorrect information.	5	11	4	3.95	3
3	The translation correctly handles sentence complexity and syntactic structures.	3	13	4	4.05	1
4	The translation avoids unnatural literal translations when a more suitable phrase exists.	3	8	9	4.3	2
5	The translation effectively resolves potential grammatical ambiguities.	8	10	2	3.7	4

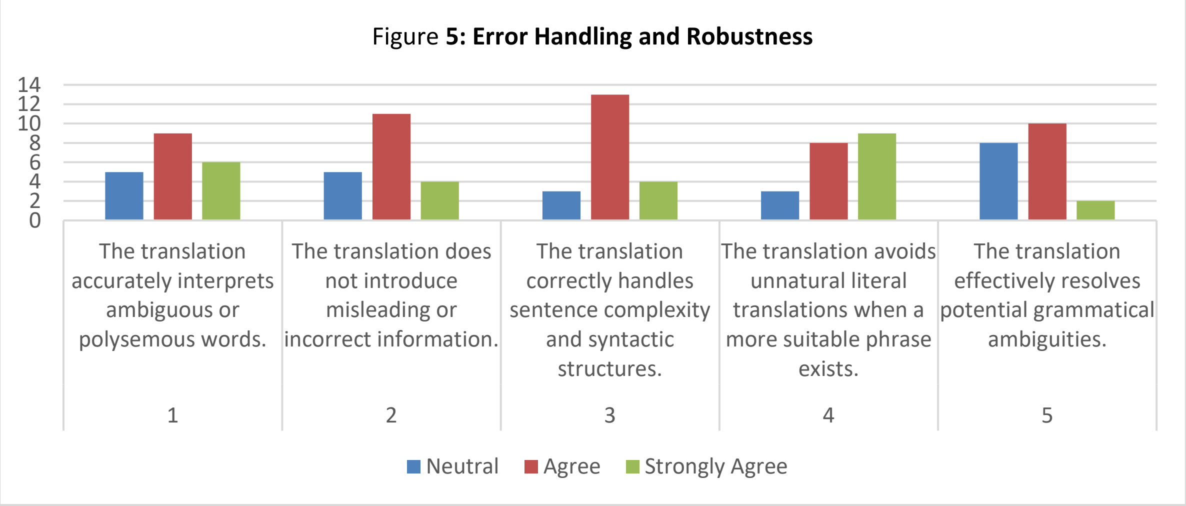
Figure **5: Error Handling and Robustness**

Table 5 and Figure 5 reveal that LLM has demonstrated such the accuracy and robustness of a translated text. The two top-ranked items, "The translation correctly handles sentence complexity and syntactic structures" and "The translation accurately interprets ambiguous or polysemous words," obtained a mean score of 4.05, showing high performance in solving linguistic complexities. The second top-ranked item, "The translation avoids unnatural literal translations when a more suitable phrase exists," obtained a mean score of 4.30, which shows excellent performance in creating natural and idiomatic phrasing. The lowest-ranked item, "The translation effectively resolves potential grammatical ambiguities," had a mean of 3.70, reflecting the need for improvement in resolving grammatical complexities.

Table 6: Comparative Evaluation and Reliability

No	Item	Neutral	Agree	Strongly Agree	Mean	Rank
1	The translation is comparable to that of a professional human translator.	4	12	4	4	4
2	The translation is suitable for professional or academic use.	6	10	4	3.9	5
	The translation is more fluent than previous LLM-generated translations.	7	9	4	3.85	6
3	The translation is more accurate and fluent than previous LLM-generated translations.	4	9	7	4.15	3
4	The translation requires minimal post-editing or correction.	15	4	1	3.3	7
5	I trust this translation for official communication.		6	14	4.7	1
6	I trust this translation for formal communication		10	10	4.5	2

Figure 6 : Comparative Evaluation and Reliability

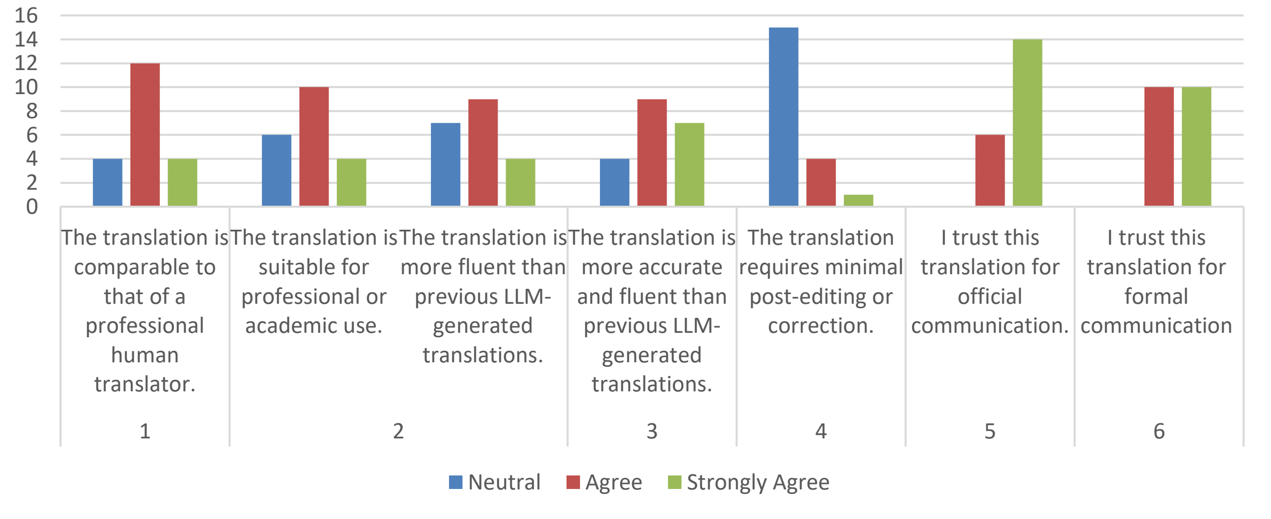

Table 5 and Figure 5 demonstrate that the respondents believe that the LLMs can perform well in handling the errors and robustness in a translated text. The highest-ranked statement, "I trust this translation for official communication," scored a mean of 4.70, indicating strong confidence in the translation's reliability for critical contexts. The second-ranked item, "I trust this translation for official communication," and for formal communication scored 4.7 and 4.50 respectively, further reinforcing the translation's reliability. LLMS can handle sentence complexity and syntactic structures," scored a mean of 4.05, indicating strong performance in managing linguistic challenges. The second-ranked statement, "The translation avoids unnatural literal translations when a more suitable phrase exists," scored a mean of 4.30, reflecting excellent LLM models' performance in producing natural and idiomatic phrasing. The lowest-ranked items, "The translation is suitable for professional or academic use," and, "The translation is more fluent than previous LLM-generated translations," scored 3.90 and 3.85, respectively, indicating positive evaluations but with notable neutral responses, highlighting enduring uncertainties about its suitability for high-stakes contexts.

7. DISCUSSION ON FINDINGS

The results of this research give significant insights into the strengths and weaknesses of LLM-translated output, from the perspective of participants. Below, we address the implications of these findings within the framework of translation accuracy, fluency, naturalness, consistency of terminology, error management, and comparative reliability.

They usually concurred that the translations also reflected the substance of the source text, indicating a key success factor for preserving semantic integrity through LLM operation. Nevertheless, the lower figures in the translation of idiomatic expressions and the use of figurative language imply potential difficulty for LLMs when dealing with cultural or contextually specific phrases. This is in line with current research, which points to the difficulties in translating idiom expressions while maintaining their intended meaning or cultural connotation. Improvements can be made in the future by developing the capacity of the model to identify and adjust for such linguistic aspects. The high ratings for grammatical correctness demonstrate that LLMs excel in producing translations that are grammatically sound and easy to follow. This is a significant achievement, as fluency is often a distinguishing factor between humans and LLMs.

Further, the participants rated the translations highly for naturalness, with many agreeing that the text read as if written by a native speaker. This is a promising result, as it suggests that LLMs are becoming increasingly capable of producing human-like translations. However, the slightly lower scores for avoiding robotic or overly literal phrasing highlight a persistent challenge in balancing literal accuracy with natural expression. This finding underscores the importance of incorporating contextual and pragmatic knowledge into translation models. Nevertheless, the consistently high ratings in this category indicate that LLMs are highly effective at using appropriate and uniform terminology. This is particularly important in specialized domains, where inconsistent or incorrect terminology can lead to misunderstandings. The strong performance in this area suggests that LLMs are well-suited for technical or domain-specific translations, provided they are trained on relevant datasets.

Participants were highly satisfied with the translations' ability to handle errors and ambiguities. This is a notable strength, as it demonstrates the model's ability to interpret ambiguous or polysemous words accurately and resolve grammatical ambiguities effectively. However, the occasional introduction of misleading information suggests that LLMs may still benefit from additional safeguards to prevent errors in critical contexts.

While participants found the translations comparable to those produced by professional human translators, there was some hesitation regarding their suitability for formal or official communication. This reflects a broader concern about the reliability of machine-generated translations in high-stakes scenarios, where even minor errors can have significant consequences. The need for minimal post-editing further underscores the importance of developing LLMs that can produce publication-ready translations with little to no human intervention.

The wide range of overall satisfaction scores highlights the diversity of participant opinions. While most participants were satisfied with the translations, a small subset expressed significant dissatisfaction, particularly regarding the translations' suitability for formal contexts. This divergence in opinions suggests that LLM-generated translations may not yet meet the expectations of all users, particularly those with high standards for accuracy and reliability.

8. CONCLUSION

The results of this research yield important information on the strengths and weaknesses of LLM-translated outputs, according to participants. Below, the researcher explains the implications of the findings in the light of translation accuracy, fluency, naturalness, consistency of terminology, error management, and comparative reliability, compared with earlier research. Most of the participants concurred that the translations retained the meaning of the original text, which is a key measure of the LLM's semantic integrity. The fact that even the lower scores for processing idiomatic expressions and figurative language still reflected a high level of semantic preservation indicates that LLMs are highly effective at maintaining semantic integrity. Future developments may emphasize how the model can better detect and learn from such linguistic characteristics, by using culturally rich datasets or context-sensitive algorithms. The high scores for grammatical correctness and sentence flow show that LLMs are very good at generating translations that are grammatically correct and readable. This is a notable accomplishment because fluency is the most important differentiator between human and machine translations. Yet, occasional instances of stilted phrasing, commented on by participants. Participants scored the translations very highly for naturalness, with most concurring that the text sounded as though it had been written by a native speaker. This is an encouraging finding since it indicates that LLMs are more and more able to generate human-like translations. However the slightly reduced scores on evading robotic or overly literal phrasing reveal an ongoing problem of reconciling literal accuracy with natural expression. The uniformly high ratings in this category suggest that LLMs excel at applying appropriate and consistent terminology. This is especially significant in technical domains, where inconsistent or inaccurate terminology may cause confusion. This also highlights the requirement for specialized training and tuning within domains to secure terminological accuracy in various areas. The participants were pleased with the translations' capacity to manage errors and ambiguities, especially in translating polysemous words and untangling grammatical intricacies. This is a significant advantage, as it indicates the model's capacity to cope with linguistic complexities well. But the sporadic injection of misleading information, which some participants pointed out, implies that LLMs might still be aided by further checks to avoid errors in high-stakes situations.

Though the participants found the translations to be equal to those done by professional human translators, there was some reservation towards their appropriateness for official or formal communication. This is in line with a general concern regarding the stability of machine translation in high-risk situations, as noted by Baker (2018). The requirement for minimal post-editing, as reflected in the lower scores in this category, further highlights the necessity of creating LLMs that can generate publication-ready translations with minimal or no human touch. The broad variation in overall satisfaction scores reflects the range of opinions among participants. Although most participants were satisfied with the translations, there was a small minority that voiced strong dissatisfaction, especially about the appropriateness of the translations for formal situations. This difference in opinion indicates that LLM-translated text may not yet be satisfying all users, especially those with strict standards for precision and trustworthiness. The results of this research have a number of implications for the machine translation and natural language processing fields. To close the gap between human and machine translations the excellent performance of LLMs in fluency, term consistency, and error management indicates that they are becoming more capable of closing the gap between human and machine translations. The issues related to idiomatic expressions, naturalness, and reliability indicate that there is still a need for research and development.

The excellent ratings in terms of terminology and consistency imply that LLMs are ideal for domain-specific usage, including legal, medical, or technical translation. The development of customized models for these topics, as well as importing external knowledge bases to enhance performance, is a direction that future research might pursue.

8.1.Recommendations

This study provides a comprehensive evaluation of LLM-generated translations, highlighting their strengths in fluency, terminology consistency, and error handling while also identifying areas for improvement, such as handling idiomatic expressions and building trust in formal contexts. The findings underscore the potential of LLMs to revolutionize the field of translation, offering high-quality, human-like translations that are increasingly comparable to those produced by professional human translators. However, the study also emphasizes the need for continued research and development to address persistent limitations, particularly in preserving cultural nuances, resolving ambiguities, and ensuring reliability in high-stakes scenarios. By focusing on enhancing contextual understanding, improving naturalness, and building trust through mechanisms such as error detection and user feedback, future iterations of LLMs could further bridge the gap between machine and human translations. This would not only expand their applicability across professional, academic, and official domains but also make high-quality, accessible language services available to a global audience, transforming the way we communicate across languages. By focusing on enhancing contextual understanding, improving naturalness, and building trust, future iterations of LLMs could further bridge the gap between machine and human translations, making them a viable option for a wider range of applications. The hesitation to use LLM-generated translations in formal or official communication highlights the need for greater transparency and reliability. The wide range of overall satisfaction scores suggests that LLMs may need to be tailored to different user needs and expectations. Offering customizable translation settings or allowing users to provide feedback on translations could help improve user satisfaction.

REFERENCES

Al-Kaabi, M. H., AlQbailat, N. M., Badah, A., Ismail, I. A., & Hicham, K. B. (2024). Examining the Cultural Connotations in Human and Machine Translations: A Corpus Study of Naguib Mahfouz's Zuqāq al-Midaqq. Journal of Language Teaching and Research, 15(3), 707-718.

Alves, D. M., Rei, R., Farinha, A., de Souza, J. G. C., & Martins, A. F. T. (2022). Robust MT evaluation with sentence-level multilingual augmentation. Conference on Machine Translation, (pp. 469-478). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.43/

Al-Wasy, B. Q., & Mohammed, O. S. M. (2024). Strategies of Translating Euphemistic Expressions from Arabic into English: A Comparative Study of Artificial Intelligence Models with Human Translation. *Humanities and Educational Sciences Journal*, (40), 826–855. https://doi.org/10.55074/hesj.vi40.1121

Anari, S. M. (2004). Accuracy, clarity, and naturalness in the translation of religious texts. Iranian Journal of Translation Studies, 2(5). https://journal.translationstudies.ir/ts/article/view/28

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Baker, M. (2018). In Other Words: A Coursebook on Translation. Routledge.

Barrault, L., Bojar, O., Costa-jussà, M. R., Gal, L., & Graham, Y. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). Proceedings of the Fourth Conference on Machine Translation.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: A case study. Computer Speech & Language, 49, 52-70.

Dahia, I., & Belbacha, M. (2024). Machine-Learning-based English Quranic Translation: An Evaluation of ChatGPT. International Journal of Linguistics, Literature & Translation, 7(8).

Davoudi Sharifabad, E., & Rajabi Fakhrabadi, F. (2023). Assessing the Quality of Hidden Proverbs Translation in the Holy Qur'ān: Human vs. Artificial Intelligence English Translations. International Journal of Textual and Translation Analysis in Islamic Studies, 1(4), 351-367.

Denecke, K., May, R., LLM HealthGroup, & Romero, O. R. (2024). Potential of large language models in health care: Delphi study. Journal of Medical Internet Research, 26, e52399.

Falempin, A., & Ranadireksa, D. (2024, December). Human vs. Machine: The Future of Translation in an AI-Driven World. In Widyatama International Conference on Engineering 2024 (WICOENG 2024) (pp. 177-183). Atlantis Press.

Forcada, M. L. (2017). Making sense of neural machine translation. Translation Spaces, 6(2), 291-309.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications, and evaluation. Journal of Artificial Intelligence Research, 61, 65-170.

Graham, Y., Haddow, B., & Koehn, P. (2019). Translationese in machine translation evaluation. arXiv preprint arXiv:1906.09833.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints, 3.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. arXiv preprint arXiv:1803.05567.

Isabelle, P., Cherry, C., & Foster, G. (2017). A challenge-set approach to evaluating machine translation. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

Iyer, V., Chen, P., & Birch, A. (2023). Towards effective disambiguation for machine translation with large language models. Conference on Machine Translation. arXiv preprint arXiv:2309.11668.

Jermakowicz, E. K. (2023). The coming transformative impact of large language models and artificial intelligence on global business and education. Journal of Global Awareness, 4(2), 1-22.

Jiang, Z., Lv, Q., Zhang, Z., & Lei, L. (2024). Convergences and Divergences between Automatic Assessment and Human Evaluation: Insights from Comparing ChatGPT-Generated Translation and Neural Machine Translation. arXiv preprint arXiv:2401.05176.

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6, https://doi.org/10.1016/j.nlp.2023.100048

Koehn, P. (2020). Neural machine translation. Cambridge University Press.

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation.

Lankford, S. (2024). Enhancing Neural Machine Translation of Low-Resource Languages: Corpus Development, Human Evaluation and Explainable AI Architectures. arXiv preprint arXiv:2403.01580. https://doi.org/10.48550/arXiv.2403.01580

Läubli, S., Sennrich, R., & Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. arXiv preprint arXiv:1808.07048.

Lyu, C., Du, Z., Xu, J., Duan, Y., Wu, M., Lynn, T., ... & Wang, L. (2023). A paradigm shift: The future of machine translation lies with large language models. arXiv preprint arXiv:2305.01181.

Mahdy, O. S. M.., Samad, S. S., & Mahdi, H. S. (2020). The attitudes of professional translators and translation students towards computer-assisted translation tools in Yemen. *Journal of Language and Linguistic Studies*, *16*(2), 1084-1095.

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Mohamed, Y. A., Khanan, A., Bashir, M., Mohamed, A. H. H., Adiel, M. A., & Elsadig, M. A. (2024). The impact of artificial intelligence on language translation: a review. Ieee Access, 12, 25553-25579.

Mohsen, M. (2024). Artificial intelligence in academic translation: A comparative study of large language models and google translate. PSYCHOLINGUISTICS, 35(2), 134-156. https://doi.org/10.1080/23311886.2024.2410998

Muñoz Andrés, P. (2024). The Impact of ISO Certifications, Machine Translation (MT), & Large Language Models (LLMs) in the Quality of English into Spanish Translations. http://hdl.handle.net/10366/163499

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. Iscience, 27(10).

Nazeer, I., Khan, N. M., Nawaz, A., & Rehman, J. (2024). An Experimental Analysis of Pragmatic Competence in Human-ChatGPT Conversations. Pakistan Journal of Humanities and Social Sciences, 12(1), 424-435.

Nordin, S., & Schmidt, A. (2024). Optimizing ChatGPT for Enhanced Machine Translation: A Systematic Approach to Contextual Accuracy and Cross-Lingual Consistency. Baltic Multidisciplinary Journal, 2(2), 155-160. https://doi.org/10.5281/

Olmedilla, M., Romero, J. C., Martínez-Torres, R., Toral, S., & Galvan, N. R. (2024). Evaluating coherence in AI-generated text.

Orlando, M., Liao, S., & Kruger, J-L. (2024). Translation and Interpreting technologies and their impact on the industry, A report prepared for the National Accreditation Authority for Translators and Interpreters (NAATI). Macquarie University

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Qamar, U., & Raza, M. S. (2024). Machine Translation Using Deep Learning. In Applied Text Mining (pp. 449-494). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-51917-8_12

Qian, M., Newton, C., & Qian, D. (2021). Cultural understanding using In-context learning and masked language modeling. In HCI International 2021-Late Breaking Papers: Multimodality, extended Reality, and Artificial Intelligence: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings 23 (pp. 500-508). Springer International Publishing. https://doi.org/10.1007/978-3-030-90963-5_38

Raunak, V., Menezes, A., Post, M., & Awadallah, H. H. (2023). Do GPTs produce less literal translations? ArXiv, abs/2305.16806.

Richards, E., & Martinez, I. (2024). Exploring Contextual Understanding in Large Language Models: A New Era of NLP. Eastern European Journal for Multidisciplinary Research, 3(2), 1-6. http://snmzpublisher.com/index.php/eejmr/article/view/36

Specia, L., Turchi, M., Federico, M., & Negri, M. (2018). Machine Translation Evaluation: Recent Trends and Future Outlook. Proceedings of the NAACL-HLT 2018.

Toral, A., Wieling, M., & Way, A. (2018). Post-editing effort of a novel with statistical and neural machine translation. Frontiers in Digital Humanities, 5(9).

Yadav, B. (2024). Generative AI in the Era of Transformers: Revolutionizing Natural Language Processing with LLMs. J. Image Process. Intell. Remote Sens, 4(2), 54-61. https://doi.org/10.55529/jipirs.42.54.61