ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 14, No. 4, pp. 985-994, April 2024
DOI: https://doi.org/10.17507/tpls.1404.07

An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score

Mozhgan Ghassemiazghandi
School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang, Malaysia

Abstract—**Traditional views have long held that machine translation cannot achieve the quality and accuracy of human translators, especially in complex language pairs like Persian and English. This study challenges this perspective by demonstrating that ChatGPT-4, with access to vast amounts of multilingual data and leveraging advanced large language model algorithms, significantly outperforms widely utilized open-source machine translation tools and approaches the realm of human translation quality. This research aims to critically assess the translation accuracy of ChatGPT-4 against a traditional open-source machine translation tool from Persian to English, highlighting the advancements in artificial intelligence-driven translation technologies. Using Bilingual Evaluation Understudy scores for a comprehensive evaluation, this study compares the translation outputs from ChatGPT-4 with MateCat, providing a quantitative basis for comparing their accuracy and quality. ChatGPT-4 achieves a BLUE score of 0.88 and an accuracy of 0.68, demonstrating superior performance compared to MateCat, with a 0.82 BLUE score and 0.49 accuracy. The results indicate that the translations generated by ChatGPT-4 surpass those produced by MateCat and nearly mirror the quality of human translations. The evaluation demonstrates the effectiveness of OpenAI's large language model algorithms in improving translation accuracy.**

Index Terms—**BLEU score evaluation, ChatGPT-4 translation, large language models, machine translation accuracy, translation quality assessment**

I. Introduction

Global communication and understanding are crucial in this information era because of linguistic barriers. An automated translation technology named Machine Translation (MT) paved the way for bridging the gap between different languages worldwide. Rawling and Wilson (2021) stated that MT facilitates information exchange by translating text from one language to another. With the need for effective cross-cultural communication, developing MT systems that can translate accurately has become a key focus of research, especially in language and translation (Wu et al., 2016). One of the most significant developments widely utilized today is ChatGPT, an advanced Large Language Model (LLM) series developed by OpenAI. This technology demonstrated an excellent performance in comprehending human context and generating text mimicking human language (Liu et al., 2023). This Artificial Intelligence (AI) model leverages a large amount of training data and advanced algorithms to capture underlying patterns, semantics, and languages. The ability of ChatGPT models, specifically ChatGPT-4, to understand and generate human-like responses in any language has transformed the domain of Natural Language Processing (NLP), which provides an avenue for developing MT systems that will outperform the MT traditional methods (Adedokun et al., 2023). The transition from Rule-Based Machine Translation (RBMT) methods to data-driven MT marked remarkable progress in language and translation (Stahlberg, 2020). Although the advancement in MT is promising because of the accessibility of different translation automation, it is essential to evaluate its performance to ensure quality standards for the translation it generates. Han (2022) emphasized that MT evaluation of statistical and neural machine translation is vital to assess the credibility and constraints of different available MT systems.

This study aims to evaluate the translation quality of ChatGPT translation from Persian to English by utilizing automated metrics such as Bilingual Evaluation Understudy (BLEU) and accuracy. BLEU is a preeminent statistical metric that quantifies the similarity between machine translations and their human-generated counterparts (Reiter, 2018). A web-based translation tool, MateCat, will be used as the comparative baseline to give insight into which MT tool performs better in translating complex language pairs like Persian and English. The study hypothesizes that the translation quality of ChatGPT-4 outperforms MateCat and is close to the reference translation because ChatGPT-4 utilizes advanced large language modeling. This study will utilize a Persian history book as the source text and its human-generated translation as the reference text. The translations from ChatGPT and MateCat will be evaluated to measure the BLEU score and accuracy of the translations from Persian to English. All the text undergoes data preprocessing to ensure uniformity and consistency among the corpus and will generate a reliable MT evaluation result. This study contributes to the advancements of MT by providing comprehensive insights about the results, supporting automatic metrics with human evaluation, and delivering knowledge of the strengths and weaknesses of ChatGPT-4, the Persian-to-English translation setting.

II. Literature Review

The advent of computers and AI heralded an advancement in translation with the development of MT systems (Poibeau, 2017). MT has the most significant potential to overcome language barriers and facilitate cross-lingual communication (Rivera-Trigueros, 2022) despite generating more nuanced findings on variations (Sakamoto, 2020). With this, research and advances in MTs have been expanding rapidly for several decades (Wang et al., 2022).

Evolution of Machine Translation

The first concept of MT appears in Warren Weaver's (1955) Memorandum on Translation, and from then on, the development of MT technologies arose, which resulted in the initial non-numerical usage of computers (Kenny, 2022). Wang (2024) remarked that the early approaches of MT were rule-based systems that relied on dictionaries, grammar, and transfer rules to generate translations. The study of Bhadwal et al. (2020) utilized Rule-Based Machine Translation (RBMT) to translate prominent language features of Hindi and Sanskrit, effectively addressing the challenge of polysemy in verb translation. While RBMT could produce high-quality translation, these systems needed improvement in handling ambiguity and idiosyncrasies of new language pairs or domains (Harsha et al., 2022; De Martino et al., 2023). The advent of statistical methods, notably with IBM's Candide system and Google Translate, marked a significant advancement of MT from RBMT (Jumanto et al., 2022). Statistical Machine Translation (SMT) models utilized large corpora of bilingual texts to infer translation probabilities, improving the quality of translations (Kahlon & Singh, 2023). Abidin and Ahmad (2021) built an SMT of Indonesian to the Lampung Nyo dialect, achieving 45.26% accuracy for 3,000 sentences, which states that SMT will improve the results over time when trained with large datasets. However, Mishra (2024) states that SMT has constraints on long-range dependencies and complex linguistics due to the scarcity of large parallel corpora. With this, Neural Machine Translation (NMT) marked a groundbreaking paradigm shift in MT as it leverages a large amount of data trained in deep learning (DL) to generate translation in different languages (Amin & Mandapuram, 2021). Zaghlool and Khasawneh (2023) emphasized the significance of MT as an accessible and efficient tool in translation. Moreover, the emergence of AI, like ChatGPT, paved the way for proficient MT tools enabling translations between various languages (Sanz-Valdivieso & López-Arroyo, 2023).

Neural Machine Translation

NMT leverages neural networks of a DL algorithm to model text sequences and address low-resource language pair issues in MT (Ranathunga et al., 2023). The process consists of two subnetworks, an encoder, and a decoder, communicating together, which is the foundation of the NMT models (Mohamed et al., 2021). Forcada and Ñeco (1997) developed a simple translation task that uses two feed-forward neural networks where the machine acquires an internal representation of the input, and the machine will decode it to generate a translated text.

For instance, Cho et al. (2014) employed two recurrent neural networks (RNNs) to encode a source sentence of varying length into a fixed-length vector and then decode the vector back into a target sentence of varying length. Bahdanau et al. (2016) proposed a stacked RNN with an attention model to convert the source sentence into a continuous vector representation to address the fixed vector problem. However, Sutskever et al. (2014) asserted that NMT that utilizes RNN with Long Short-Term Memory (LSTM) units delivers a performance almost as good as the conventional phrase-based MT system when translating from English to French. LSTM-based neural network architecture was improved significantly for NMT over traditional RNN-based architecture (Olah, 2015). The Transformer architecture, as proposed by Vaswani et al. (2017), revolutionized NMT by replacing RNNs with self-attention mechanisms, facilitating parallelization during training, enabling faster convergence, and improved performance. Recent developments in NMT architectures have proved superior performances in different languages, with immense potential for further advancements.

Machine Translation Evaluation

Evaluating MT presents significant challenges due to the large and unknown corpus, which hampers the precision of automated metrics. Consequently, extensive human evaluation is required to provide a dependable benchmark for assessing the quality and advancement of MT (Freitag et al., 2021). Even human translations are possibly biased and subjective, considering the possibility of several translations for an original text that could be deemed accurate (Rivera-Trigueros, 2022). Significant grey areas still require attention, particularly ambiguity and the semantic complexities inherent in poetic expressions crucial for cross-cultural and multilingual literary translation (Ghassemiazghandi, 2023; Fakih et al., 2024). Automated metrics evaluate the results of an MT system in reference to one or more human-generated translations (Han, 2016). Initially, Levenshtein (1966) developed a Word Error Rate (WER) where the translation quality is based on edit distance, counting equal weight for substitutions, deletions, and insertions without considering word reordering. This method reflects the inaccuracy of translations where word order differs significantly between the output and reference translations. Position-Independent Word Error Rate (PER) and Translation Error Rate (TER) address this issue, focusing on word comparisons without considering order and penalizing word reordering, respectively (Tillmann et al., 1997; Snover et al., 2006). With further advancements in MT evaluation, Papineni et al. (2002) developed Bilingual Evaluation Understudy (BLEU) metrics to quantify the translation quality of MT by aggregating the total number of words and phrases (n-grams) that are shared across machine and reference translations. The metric imposes a penalty for excessively brief translations. Although BLEU correlates with human translations (Kocmi et al., 2021), the ratio of matched n-grams to the total number of n-grams in the reference translation is not taken into account (Maruf et al., 2021). However, BLEU does not capture fluency, semantic similarity, or word order variations and can penalize correct translations with different phrasing (Segonne & Mickus, 2023; Haque et al., 2022). Despite its limitation, Rivera-Trigueros's (2022) systematic review reveals that BLEU is the most used automatic metric, which is also evident in Marie et al. (2021) meta-evaluation of 769 papers. While BLEU serves as a benchmark, researchers actively explore alternative metrics and integrate human evaluation for a more comprehensive assessment (Evtikhiev et al., 2023; Freitag et al., 2022). Even if several studies (Chatzikoumi, 2019; Way, 2018) claimed that integrating human evaluation and automated MT metrics will obtain the most dependable outcome, only 22% of works analyzed in the study of Marie et al. (2021) employed this combined method, which indicates a lack of study and development in the domain of MT that focuses on translation and language. By addressing these areas, researchers can better understand MT quality, paving the way for advanced and human-like MT systems.

ChatGPT Machine Translation

Recent advancements in AI, particularly the advanced LLM techniques, have brought remarkable improvements in building more effective translation systems (Chowdhery et al., 2022). Generative Pre-trained Transformer (GPT), a renowned advanced language model created by OpenAI, has acquired considerable scrutiny for its capacity to comprehend and produce coherent and logical text (Hendy et al., 2023; Sahari et al., 2024). However, the translation generated needs thorough assessment because of the need for more understanding of domain terminologies and the cultural context of the model (Khoshafah, 2023). Jiao et al. (2023) revealed that GPT-3 outperforms Google Translate, an MT tool, in multilingual translation prompts and robustness for European languages. Another study by Banat and Abu Adla (2023) demonstrates that, although GPT-3 translates Arabic text to English with high accuracy, it requires post-editing to adequately capture cultural context. The findings of each study support Hendy et al. (2023) findings that GPT models have constraints when the language has scarce resources. With this, Chowdhery et al. (2022) emphasize the importance of thoroughly evaluating the translation produced by MT rather than solely relying on automated metrics.

III. Methodology

This section will discuss the methods for evaluating the translation quality of two MT systems in Persian-to-English with reference to human-generated English translation. Figure 1 illustrates the process of this study, which is divided into several key sections: 'Data Source,' discussing the data used; 'Data Preparation,' explaining the two MT systems employed; 'Data Preprocessing,' describing how the data was cleaned; and 'Evaluation Metrics,' specifying the methods used to evaluate the MT translations, along with 'Experimental Setup'.

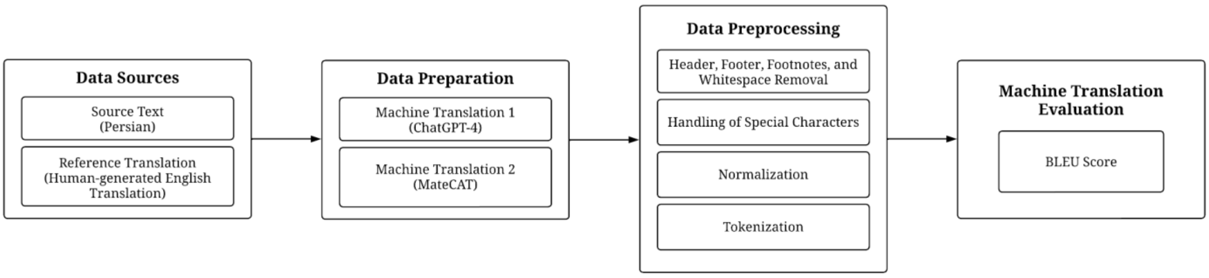

Data Sources
Source Text (Persian)
Reference Translation (Human-generated English Translation)
Data Preparation
Machine Translation 1 (ChatGPT-4)
Machine Translation 2 (MateCAT)
Data Preprocessing
Header, Footer, Footnotes, and Whitespace Removal
Handling of Special Characters
Normalization
Tokenization
Machine Translation Evaluation
BLEU Score

Figure 1. Design of the Study

A. Data Source

The book Amiran e Darbar, authored by Akbar Tehrani Shafagh in 2006, is selected as the dataset for the source text of this study. "Princes of the Court: Memoirs of the Seil Sepor Family," translated into English in 2023 by the same author, is used as a translation reference or human transcription. "Amiran e Darbar" is a memoir that chronicles the history of the Seil Sepor family in a narrative form that intersects with the pivotal and historical events of Iran while the events of this family are going on. What makes this book particularly intriguing is the author's background as a legal expert, which infuses the narrative and descriptions of social and cultural events with a complex blend of literary and legal vocabulary. These attributes enhance the book's appeal and add a layer of complexity to its text, making the translation process both challenging and rewarding. The choice of this book was further motivated by its rich cultural narratives and historical terms, demanding linguistic proficiency and verbal finesse for an appropriate translation. Even for seasoned translators, navigating the novel vocabulary, specific terminologies, and the need to preserve the cultural and social nuances of the text presents a considerable challenge. Thus, despite advancements in AI and MT algorithms, translating such a text remains daunting, laden with difficulties in capturing its intricate and rich cultural essence.

B. Data Preparation

The English translation of the Persian history book has been executed by utilizing two MT systems: ChatGPT-4 and MateCat. ChatGPT-4 is a promising translation tool that leverages advanced large language model algorithms to translate text between multiple languages (Castillo-Gonz ález et al., 2022). MateCat is also a commonly used MT tool that utilizes Google Translate as its MT engine. Its popularity stems from its availability and the comprehensive reports it provides (Quintana & Castilho, 2022).

C. Data Preprocessing

Kang et al. (2021) asserted that normalizing text inputs is crucial in MT as it eliminates nuances and ensures consistent structure among data. With this, data cleaning is applied to the data sources to ensure uniformity and is ready for evaluation. The process involves removing page numbers, headers, footers, and endnotes in the reference English translation. Removing whitespaces and handling special characters and line endings in each data is applied to ensure compatibility among each translation. Lastly, the data will undergo tokenization to identify the sentence and word count essential for data exploration and translation quality assessment.

D. Evaluation Metric

The Bilingual Evaluation Understudy (BLEU) metric is utilized to evaluate the similarity between the English human-generated text, also known as reference text, and MT based on matching words and phrases, known as n-grams (Papineni et al., 2002). Equation 1 shows how the BLEU functions work behind the system where r is the output length divided by c, the reference length, which implements the brevity penalty for short translations. The result will be multiplied by the geometric average precision where n-grams are size 4.

$$\log BLEU \ = \min\left(1 - \frac{r}{c}, 0\right) \left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}} \tag{1}$$

The *corpus_bleu* function in the Natural Language Toolkit (NLTK) library allows automation in quantifying translation quality. The BLEU score represents a value between 0 and 1, where 1 highlights the exact correspondence between the reference text and MT. To further extend the translation quality evaluation, the *accuracy* (eq 2) of the proportion of tokens between reference translation and MT is measured to identify lexical overlap.

$$Accuracy = \frac{Total\ no.of\ overlapping\ words\ in\ the\ reference\ and\ MT}{Total\ unique\ words\ in\ the\ reference} \times 100 \tag{2}$$

Figure 2 shows the entire process of how the data sources are utilized to perform automatic metric MT evaluation in this study.

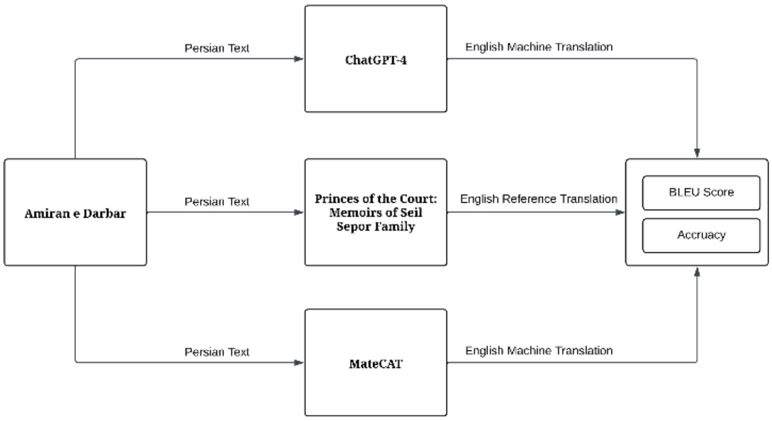

Persian Text
ChatGPT-4
English Machine Translation
Amiran e Darbar
Persian Text
Princes of the Court: Memoirs of Seil Sepor Family
English Reference Translation
BLEU Score
Accruacy
Persian Text
MateCAT
English Machine Translation

Figure 2. BLEU Evaluation Process

E. Experimental Setup

This study utilizes Pycharm to perform automatic metrics used to quantify translation quality. The program provides a flexible environment for implementing processes like calling data, data preprocessing, and necessary functions to calculate BLEU scores and accuracy.

IV. Results

The results section of this study presents a detailed analysis of the performance of ChatGPT-4 from Persian-to-English translation. The findings are organized into three main sections to provide a comprehensive analysis of our findings.

A. Data Exploration

After tokenizing the data sources and translations produced by the MT systems, Table 1 shows that the word and sentence counts are close in number and do not differ by over 1k words and over 200 sentences, suggesting comparable lengths over corpora.

TABLE 1
DATA SUMMARY

Corpus	Total No. of Words	Total No. of Sentences
Source Text	9230	181
Reference Translation	9909	370
ChatGPT-4	9280	262

The source text, which contains the Persian language, is the basis of the machine translations of ChatGPT-4 and MateCat. The results show that ChatGPT-4 has a closer word count than the source text compared to MateCat, indicating preserving the original content. However, it has a higher sentence count than MateCat, which is still higher than the source text, indicating more fragmented sentences than the source text.

B. BLEU Scores

ChatGPT-4 achieved a noteworthy result, demonstrating its effectiveness in translating Persian text to English in reference to human-generated English translation. Table 2 illustrates that the ChatGPT-4 English translation outperforms the open-source MT tool MateCat. Notably, the BLEU score, a measure of translation quality, is significantly higher for ChatGPT-4(0.88) compared to MateCat (0.82).

TABLE 2
BLUE SCORE

Machine Translation Tool	BLEU Score
ChatGPT-4	0.88
MateCat	0.82

The BLEU score of ChatGPT-4 is remarkably high, suggesting that the translations produced by this MT closely align with the human-generated English translation.

C. Accuracy

In addition to BLEU evaluation, the focus extends to identifying accuracy to highlight the proper alignment between machine-translated text and reference translations over the corpus. Table 3 shows that ChatGPT-4 has higher accuracy (0.68) than MateCat (0.49).

TABLE 3
ACCURACY

Machine Translation Tool	Accuracy
ChatGPT-4	0.68
MateCat	0.49

Although ChatGPT-4 achieved a higher BLEU score and accuracy score than MateCat, our study also highlights the impact of specific challenges, such as dates, names, idioms, cultural items, and numerical expressions.

V. DISCUSSION

This research demonstrates significant advancements in MT, especially from Persian to English. ChatGPT-4 has a higher BLEU score and accuracy score than MateCat, indicating that ChatGPT-4 produces translation closer to the reference text than MateCat. Jiao et al. (2023) asserted that ChatGPT-4 is versatile for various text translation tasks. It addresses the weakness of Google Translate, the MT engine used by MateCat (Quintana & Castilho, 2022), in translating low-frequency words. Their study also supports that ChatGPT-4 is better as it predicts abbreviations better. These findings align with our objective of demonstrating the effectiveness of ChatGPT-4 over MateCat in terms of translation quality. The result shows the relevance of utilizing MT as a practical tool for translation tasks. However, ChatGPT-4 indicates potential for more diverse translations with different lexical choices due to its varied corpora.

The section below explores common issues encountered in Persian-English translations. The tables utilize abbreviations to distinguish between the types of translations: "ST" for Source Text, representing the original material; "HT" for Human Translation, denoting the reference corpus; and "MT" for Machine Translation, encompassing translations by ChatGPT-4 and MateCat.

TABLE 4
SEMANTIC ACCURACY

ST		آموزگاری که از وقت طلوع والشمس والضحی تا هنگام واللیل اذا سجی آرام و قرار نداشت
HT		A teacher who knew no rest from dawn 'til dusk, tirelessly teaching from 'Wash-Shams wal Duhā' to the time of 'Wal-Lail idhā Sajā' Footnote 1: it refers to ayat or verse of Ash-Shams surah which is the 91st surah of the Qur'an, "By the Glorious Morning Light". [Translator’s Note] Footnote 2: it refers to ayat or verse of Ash-Shams surah which is the 91st surah of the Qur'an, "And by the Night when it is still". [Translator’s Note]
MT	ChatGPT-4	A teacher who, from the time of sunrise and the bright morning until the moment of night when it settles, had no peace or rest.
	MateCat	A teacher who was not calm and quiet from the time of sunrise, sunset, and sunset until the night of Aza Saji

ChatGPT-4 captures the essence of the ST by conveying the idea of a teacher's restlessness from sunrise to nightfall. It does not, however, explicitly mention the Quranic references, similar to the human translation, but the context is respected. MateCat introduces inaccuracies, such as mentioning "sunset" twice and misinterpreting "والضحی" (Wal Duhā). The phrase "night of Aza Saji" is a misinterpretation, showing confusion in translating "واللیل اذا سجی" (Wal-Lail idhā Sajā). Moreover, it completely misses the Quranic reference, leading to a loss of cultural and contextual depth. After comparing the output of ChatGPT-4 and MateCat, it was observed that ChatGPT-4 provides a better translation in terms of accurately capturing the essence and tone of the original Persian text, as well as maintaining coherence and fluency without introducing inaccuracies or redundant phrases. However, both ChatGPT-4 and MateCat failed to understand the cultural and Quranic references that the human translator thoughtfully translated in the book and added footnotes for further explaining these Quranic terms to the readers and helping them to fully understand the ST while enjoying the Arabic rhythm of the 'Wash-Shams wal Duhā' and 'Wal-Lail idhā Sajā' in the ST. Comparing the output of these two MT systems, it is observed that ChatGPT-4 provides a better translation compared to MateCat in terms of accurately capturing the essence and tone of the ST and maintaining coherence and fluency without introducing inaccuracies or redundant phrases. However, ChatGPT4 and MateCat could not transfer the cultural and Quranic references that the HT thoughtfully rendered, which is essential for a thoroughly informed and culturally sensitive understanding of the ST.

TABLE 5
CULTURAL SENSITIVITY

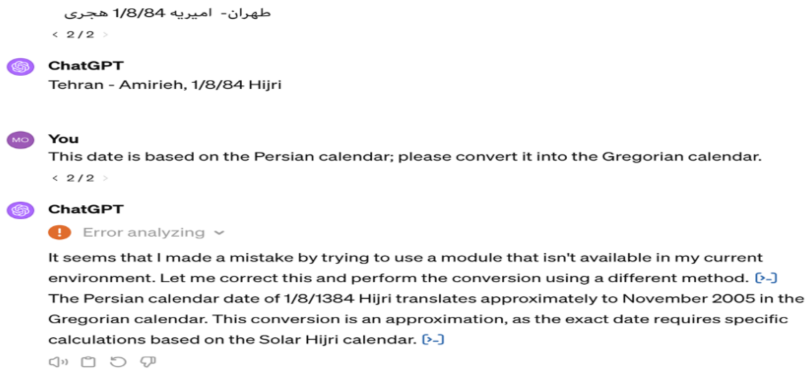
ST		ایرانیان با مراسم شب چهارشنبه سوری به ، پیشواز نوروز می روند
HT		Iranians welcome Nowruz with the Chaharshanbe Suri ceremony
MT	ChatGPT-4	Iranians welcome Nowruz with the Chaharshanbe Suri ceremony.
	MateCat	Iranians celebrate Nowruz with a ceremony on Wednesday evening.

ChatGPT-4 accurately translated چهارشنبه سوری to "Chaharshanbe Suri," shows cultural sensitivity, recognizing that AI could understand this Iranian tradition and its distinct identity. However, MateCat failed to mention "Chaharshanbe Suri," suggesting a lack of cultural sensitivity to this MT. ChatGPT -4's success in identifying this Iranian tradition and culture is due to its cultural contextualization and deep contextual knowledge provided by training data of this AI through literature, history, and cultural studies. Therefore, unlike MateCat, ChatGPT-4 could still identify certain traditions and practices in Iranian culture. This approach weakens the unique cultural significance of the event. Between ChatGPT-4 and MateCat, ChatGPT-4 provides a better MT output for this text. It matches the human translation precisely, accurately reflecting the cultural event's name and significance about Nowruz. These findings not only ensure clarity but also preserve the cultural integrity of the original Persian text. MateCat, while offering a technically correct translation, falls short in conveying the specific cultural context and significance of "Chaharshanbe Suri," thus reducing the translation's overall effectiveness and richness.

TABLE 6
TRANSLATING DATE AND LOCATION

ST		طهران- امیریه 84/8/1 هجری خورشیدی
HT		Tehran - Amirieh, 84/8/1 Solar Hijri Calendar Corresponding to September 23, 2005 - Gregorian
MT	ChatGPT-4	Prompt 1: Tehran - Amirieh, 1/8/84 Hijri Prompt 2: November 2005 in the Gregorian calendar
	MateCat	Tehran-Amiriya 1/8/84 Hijri

ChatGPT-4's first answer captures the location accurately but provides the date in a simplified format without specifying the calendar system clearly as "Solar Hijri." The second prompt, as displayed in Table 6, attempts to provide a Gregorian calendar equivalent, specifying "November 2005," which shows an effort to translate the date into the Gregorian system but lacks the exact day and misinterprets the month compared to the human translation. MateCat provides a translation that mentions the location and the date in a format similar to the ST but, like ChatGPT's first prompt, does not specify the calendar system as "Solar Hijri." It does not attempt to convert the date into the Gregorian calendar, leaving the contextual information translation incomplete. The HT not only specifies the date and location but also indicates the calendar system ("Solar Hijri Calendar") and provides the exact Gregorian calendar equivalent ("September 23, 2005"). This level of detail is crucial for understanding the exact timing of the event or record, considering the audience might not be familiar with the Solar Hijri Calendar. ChatGPT-4 and MateCat both fall short of this standard. ChatGPT-4 attempts to provide a Gregorian equivalent but inaccurately, and neither MT output specifies the calendar system as "Solar Hijri," which is crucial for clarity and understanding. In this comparison, although neither ChatGPT-4 nor MateCat fully matches the human translation's accuracy and completeness, ChatGPT-4 shows an effort to bridge the cultural and calendrical gap by attempting to provide a Gregorian calendar equivalent. However, its accuracy regarding the exact Gregorian date is off. Overall, ChatGPT-4 is slightly better in providing more context by including a Gregorian date, albeit inaccurately. However, both MTs do not fully meet the standard set by human translation in terms of accuracy, specificity, and providing a complete contextual understanding of the date across different calendar systems.



طهران- امیریه 1/8/84 هجری

‹ 2 / 2 ›

ChatGPT

Tehran - Amirieh, 1/8/84 Hijri

You

This date is based on the Persian calendar; please convert it into the Gregorian calendar.

‹ 2 / 2 ›

ChatGPT

Error analyzing

It seems that I made a mistake by trying to use a module that isn't available in my current environment. Let me correct this and perform the conversion using a different method. [>-]

The Persian calendar date of 1/8/1384 Hijri translates approximately to November 2005 in the Gregorian calendar. This conversion is an approximation, as the exact date requires specific calculations based on the Solar Hijri calendar. [>-]

Figure 3. ChatGPT Translations With Varied Prompts

TABLE 7
CONVEYANCE OF METAPHORICAL MEANING

ST		همیشه آن عالیجاه چه در حضر و چه در سفر و حتی در آنات جنگ و غیر جنگ خود را کلب آستان علـی علیه السلام می خواند.
HT		He always, whether in presence or travel, in war or peace, referred to himself as 'Kalb-e Astan Ali', peace be upon him. Footnote: The dog at the threshold of Ali: In the word "Kalb" in Arabic, it means 'dog', and here it refers to the dog of the threshold of Hazrat Ali, signifying a disciple and, metaphorically, a defender and guardian. [Translator's Note]
MT	ChatGPT-4	He always, whether in presence or on a journey, and even in times of war and peace, referred to himself as the dog of the threshold of Ali, peace be upon him.
	MateCat	Always, whether in presence or on a journey, and even during war and non-war, he calls himself the club of Astan Ali, peace be upon him.

ChatGPT-4's Translation accurately catches the essence and metaphorical meaning of the original Persian text. The translation correctly interprets "کلب آستان علـی" as "the dog of the threshold of Ali," which aligns well with the reference translation. The context of being a disciple, defender, and guardian, as explained in the HT footnote, is effectively conveyed. MateCat's translation fails to interpret the metaphorical significance of "کلب آستان علـی accurately." The translation as "the club of Astan Ali" is incorrect and suggests a misunderstanding of the term "کلب" (Kalb), which means "dog" in Arabic, not "club." MT has struggled to convey the metaphorical significance and the portrayal of the protector and supporter of the Prophet as effectively as the human translator in the book's text and its footnotes (Translator's note). However, when comparing ChatGPT-4 and MateCat, the researcher was astonished that ChatGPT-4 excelled in delivering a relatively accurate translation. A Persian translator might not be aware that the word ک لب is an Arabic word meaning "dog," yet ChatGPT-4's translation has been notably precise. This accuracy could be attributed to ChatGPT-4's access to a corpus of various languages within its algorithm, demonstrating its sophisticated linguistic capabilities.

TABLE 8
POETIC ESSENCE

ST		هرکه چون خاک نیست بر در او گرفرشته است خاک بر سر او
HT		Those who rest not as dust at his revered gate, Even if angels, must accept dust upon their crown as fate.
MT	ChatGPT-4	Whoever is not like dust at his door, Even if an angel, dust be upon his head.
	MateCat	He who is not like dust on his door, dust is on his head

Translating poetry is a challenging task, as it requires an accurate and meaningful translation while maintaining the beauty and charm of the poem. Poetic translation also needs to preserve the rhythm and rhyme, uncover hidden meanings, and convey the excitement and passion inherent in the poem. ChatGPT-4's translation successfully preserves the conditional structure of the text "Even if Angels" as well as the format of the poem, accurately reflecting the poetic nature of the poem, which has both rhythms and rhymes. While the MateCat provides an accurate direct translation, it captures only a portion of the poem's meaning, focusing on humility. However, MateCat fails to accurately transfer the message of the poem's second part, the aspect of being "the dust of the door". Comparing these MT systems, the output of ChatGPT-4 more precisely captures the conditional structure of "Even if angels" in this verse, and it preserves artistic essence and conveys an accurate translation of metaphors and poetic devices while preserving the poem's rich emotional content and beauty. ChatGPT-4's translation effort strives to maintain harmony between literary devices, especially the rhythm and rhyme, ensuring the fluency and eloquence of the poem. It is worth mentioning that even a Persian translator might struggle to effectively replicate the poem's rhythms and rhymes, highlighting that a translator and poets need different skills. The skills that ChatGPT-4 has here could help the translator to demonstrate such skills adeptly.

Overall, although the MateCat has made a significant effort to preserve the essence and beauty of the poetic devices, the ChatGPT-4 has shown a commendable ability to present a natural translation that maintains the poem's meaning while capturing its emotional depth and artistic essence of a poem.

The implication of the findings emphasizes the significant advancement of MT technology from statistical rule-based methods to data-driven methods where ChatGPT-4 outperforms MateCat in translating Persian-to-English texts. The result implies that advanced large language models leveraging DL technology better capture linguistic nuances, cultural context, and semantic meaning (Stahlberg, 2020) compared to the limited language resources used in SMT. As a result, these models produce more accurate translations in terms of context, making it ideal for translation workflows. Utilizing MT tools combined with human evaluation increases efficiency and credibility and reduces costs to generate culturally significant context. The constraints of MT tools open up future research directions in improving methodologies, architectures, and domain adaptability for language with limited resources (Farooq et al., 2021).

VI. Conclusion

The study's results enhance the understanding of MT, especially regarding the complex Persian-English language pair. By comparing the BLEU scores and accuracy of ChatGPT-4 with those of MateCat—a widely used open-source MT tool—the study demonstrates ChatGPT-4's superior translation quality. Specifically, ChatGPT-4 achieved a BLEU score of 0.88 and an accuracy rate of 0.68, surpassing MateCat's BLEU score of 0.82 and accuracy rate of 0.49. The study findings challenge the prevailing skepticism surrounding MT's ability to achieve human-like precision and underscore the transformative potential of AI-driven technologies in redefining translation practices. However, it is worth noting that semantic accuracy, cultural sensitivity, translating dates and locations, metaphorical meaning, and poetic essence are some of the factors that impact the translation quality, which poses challenges significant challenges for MT systems, as they require a deep understanding of language nuances, cultural context, and context-specific meanings. By closely mirroring the quality of human translations, ChatGPT-4 represents a pivotal advancement in the field, offering promising prospects for enhancing cross-cultural communication and understanding.

VII. Limitations and Suggestions for Future Studies

While the findings of this study offer significant insight into the performance of ChatGPT-4 in terms of MT output quality and human evaluation, it is crucial to acknowledge the limitations. The reliance on BLEU scores as the sole metric for evaluation, supported by accuracy metrics, might not capture the nuances of linguistic quality and semantic accuracy in translations. While the BLEU score can provide quantitative quality translation analysis, supporting the result with other evaluation metrics is recommended to deliver a comprehensive translation assessment, including the semantic similarity, fluency, and nuances of machine translations. It follows that the dataset used for comparison only focused on unidirectional translation from Persian to English. This limitation could impact the generalizability of the study's conclusions across all languages.

As for suggestions for further studies, the promising results of this study pave the way for a deeper investigation into how AI-assisted translation can redefine the accessibility and standards of translation practices globally. Future studies should explore the broader implications of these technologies on the translation industry, focusing on their potential to complement human skills and improve global communication. Training translators to effectively use AI tools could also leverage human and MT strengths, enhancing the quality and efficiency of translation processes. The study indicates that ChatGPT-4 performs exceptionally well; therefore, the researcher suggests that future research should explore integrating AI tools into translator training programs. This approach will facilitate accessibility and global access to information worldwide, which is a pressing need. The researcher passionately asserts, "While the world is divided by physical borders, let us ensure that language does not become another border".

References

[1] Abidin, Z., & Ahmad, I. (2021). Effect of mono corpus quantity on statistical machine translation Indonesian–Lampung dialect of nyo. In *Journal of Physics: Conference Series*, *1751*(1), 12036.

[2] Adedokun, M. J., Salami, S., Onyeali, D. C., Toheeb, B. O., Adeyoyin, D., & Afuzobugwu, K. (2023). *Transforming Smallholder Farmers Support with an AI-powered FAQbot: A Comparison of Techniques*. Retrieved March 7, 2024, from https://openreview.net/forum?id=VPl472SKaB

[3] Amin, R., & Mandapuram, M. (2021). CMS - Intelligent Machine Translation with Adaptation and AI. *ABC Journal of Advanced Research*, *10*(2), 199–206. https://doi.org/10.18034/abcjar.v10i2.693

[4] Bahdanau D, Cho K, Bengio Y. (2016). *Neural machine translation by jointly learning to align and translate.* https://doi.org/10.48550/arXiv.1409.0473

[5] Banat, M., & Abu Adla, Y. (2023). Exploring the Effectiveness of GPT-3 in Translating Specialized Religious Text from Arabic to English: A Comparative Study with Human Translation. *Journal of Translation and Language Studies*, *4*(2), 1–23. https://doi.org/10.48185/jtls.v4i2.762

[6] Bhadwal, N., Agrawal, P., & Madaan, V. (2020). A machine translation system from Hindi to Sanskrit language using rule based approach. *Scalable Computing: Practice and Experience*, *21*(3), 543–554. https://doi:10.12694/scpe.v21i3.1783

[7] Castillo-González, W., Lepez, C. O., & Bonardi, M. C. (2022). Chat GPT: a promising tool for academic editing. *Data & Metadata*, *1*, 23. https://doi:10.56294/dm202223

[8] Chatzikoumi, E. (2019). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, *26*(2), 137–161. https://doi.org/10.1017/s1351324919000469

[9] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734. https://doi.org/10.3115/v1/d14-1179

[10] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ..., Fiedel, N. (2022). *Palm: Scaling language modeling with pathways.* https://doi.org/10.48550/arXiv.2204.02311

[11] De Martino, J.M., Silva, I.R., Marques, J.C.T., Martins, A.C., Poeta, E.T., Christinele, D.S., & Campos, J.P.A.F. (2023). *Neural machine translation from text to sign language.* Univ Access Inf Soc. https://doi.org/10.1007/s10209-023-01018-6

[12] Evtikhiev, M., Bogomolov, E., Sokolov, Y., & Bryksin, T. (2023). Out of the BLEU: How should we assess quality of the Code Generation models? *Journal of Systems and Software*, *203*, 111741. https://doi.org/10.1016/j.jss.2023.111741

[13] Fakih, A., Ghassemiazghandi, M., Fakih, A. H., & Singh, M. K. (2024). Evaluation of Instagram's Neural Machine Translation for Literary Texts: An MQM-Based Analysis. *Gema Online Journal of Language Studies 213, Volume 24*(1), 1730-1732. . http://doi.org/10.17576/gema-2024-2401-13

[14] Farooq, U., Rahim, M. S. M., Sabir, N., Hussain, A., & Abid, A. (2021). Advances in machine translation for sign language: approaches, limitations, and challenges. *Neural Computing and Applications*, *33*(21), 14357–14399. https://doi.org/10.1007/s00521-021-06079-3

[15] Forcada, M. L., & Ñeco, R. P. (1997). Recursive hetero-associative memories for translation. *Lecture Notes in Computer Science*, 453–462. https://doi.org/10.1007/bfb0032504

[16] Freitag, M., Rei, R., Mathur, N., Lo, C-K., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., & Martins, A.F.T. (2022). Results of WMT22 metrics shared task: Stop using BLEU–neural metrics are better and more robust. *In Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 46-68).

[17] Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460–1474. https://doi.org/10.1162/tacl_a_00437

[18] Ghassemiazghandi, M. (2023). Machine Translation of Selected Ghazals of Hafiz from Persian into English. *Arab World English Journal for Translation and Literary Studies*, *7*(1), 220–232. https://doi.org/10.24093/awejtls/vol7no1.17

[19] Han, L. (2016). Machine translation evaluation resources and methods: A survey. *ArXiv: Computation and language.* Cornell University Library. https://doi.org/10.48550/arXiv.1605.04515

[20] Han, L. (2022). *An overview on machine translation evaluation.* https://doi.org/10.48550/arXiv.2202.11027

[21] Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022). Semantic similarity metrics for evaluating source code summarization. *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*. https://doi.org/10.1145/3524610.3527909

[22] Harsha, N. S., Kumar, C. N., Sonthi, V. K., & Amarendra, K. (2022). Lexical Ambiguity in Natural Language Processing Applications. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 1550-1555). IEEE.

[23] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita,H., Kim, Y.J., Afify, M., & Awadalla H.H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation.* https://doi.org/10.48550/arXiv.2302.09210

[24] Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). *Is ChatGPT a good translator? A preliminary study.* https://doi.org/10.48550/arXiv.2301.08745

[25] Jumanto, J., Rizal, S. S., Asmarani, R., & Sulistyorini, H. (2022). The Discrepancies of Online Translation-Machine Performances: A Mini-Test on Object Language and Metalanguage. In *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)* (pp. 27-35). IEEE.

[26] Kahlon, N.K., & Singh, W. (2023) Machine translation from text to sign language: a systematic review. *Univ Access Inf Soc*, *22*, 1–35. https://doi.org/10.1007/s10209-021-00823-1

[27] Kang, X., Zhao, Y., Zhang, J., & Zong, C. (2021). Enhancing lexical translation consistency for document-level neural machine translation. *Association for Computing Machinery*, *21*, 3. https://doi.org/10.1145/3485469

[28] Kenny, D. (2022). Human and machine translation. *Machine translation for everyone: Empowering users in the age of artificial intelligence*, *18*, 23.

[29] Khoshafah, F. (2023). *ChatGPT for Arabic-English Translation: Evaluating the Accuracy.* https://doi.org/10.21203/rs.3.rs-2814154/v1

[30] Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., & Menezes, A. (2021). *To ship or not to ship: An extensive evaluation of automatic metrics for machine translation.* https://doi.org/10.48550/arXiv.2107.10821

[31] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

[32] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of CHATGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

[33] Marie, B., Fujita, A., & Rubino, R. (2021). *Scientific credibility of machine translation research: A meta-evaluation of 769 papers.* https://doi.org/10.48550/arXiv.2106.15195

[34] Maruf, S., Saleh, F., & Haffari, G. (2021). A Survey on Document-level Neural Machine Translation. *ACM Computing Surveys*, *54*(2), 1–36. https://doi.org/10.1145/3441691

[35] Mishra, R. (2024). A Comparative Analysis of Statistical and Neural Machine Translation Models. *Integrated Journal of Science and Technology*, *1*(2), 1-3

[36] Mohamed, S. A., Elsayed, A. A., Hassan, Y. F., & Abdou, M. A. (2021). Neural machine translation: past, present, and future. *Neural Computing and Applications*, *33*(23), 15919–15931. https://doi.org/10.1007/s00521-021-06268-0

[37] Olah, C. (2015). *Understanding LSTM Networks*. Retrieved March 8, 2024, from https://colah.github.io/posts/2015-08-Understanding-LSTMs

[38] Papineni, K., Roukos, S., Ward, T., & Zhu W-J.(2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Meeting on Association for Computational Linguistics*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[39] Poibeau, T. (2017). *Machine translation*. MIT Press.

[40] Quintana, R. C., & Castilho, S. (2022). A review of the Integration of Machine Translation in CAT tools. *New Trends in Translation and Technology 2022*, 214.

[41] Ranathunga, S., Lee, E. S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, *55*(11), 1-37.

[42] Rawling, P., & Wilson, P. (2021). *The Routledge Handbook of Translation and Philosophy*. Abingdon, Oxon: Routledge, Taylor & Francis Group.

[43] Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, *44*(3), 393-401.

[44] Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Lang Resources & Evaluation*, *56*, 593–619. https://doi.org/10.1007/s10579-021-09537-5

[45] Sahari, Y., Qasem, F., Asiri, E., Alasmri, I., Assiri A., & Mahdi, H. (2024). *Translation of Figurative Language: A Comparative Study of ChatGPT and Human Translators.* https://doi.org/10.21203/rs.3.rs-3921149/v1

[46] Sakamoto, A. (2020). The value of translation in the era of automation: An examination of threats. *When Translation Goes Digital*, 231–255. https://doi:10.1007/978-3-030-51761-8_10

[47] Sanz-Valdivieso, L., & Lόpez-Arroyo, B. (2023). Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation? *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology 2023*. https://doi.org/10.26615/issn.2683-0078.2023_008

[48] Segonne, V., & Mickus, T. (2023). "Definition Modeling: To model definitions." *Generating Definitions With Little to No Semantics.* https://doi.org/10.48550/arXiv.2306.08433

[49] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas* (pp. 223–231).

[50] Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, *69*, 343-418.

[51] Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS 2014)*.

[52] Tehrani Shafagh, A. (2023). *Princes of the Court: Memoirs of the Seil Sepor Family* (A. Tehrani Shafagh, Trans.). Sahami Enteshar Company. (Original work published 2006)

[53] Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based search for statistical translation. *5th European Conference on Speech Communication and Technology* (Eurospeech 1997). https://doi.org/10.21437/eurospeech.1997-673

[54] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (2017)*: 6000–6010.

[55] Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in machine translation. *Engineering*, *18*, 143-153.

[56] Wang, Y. (2024). Research of types and current state of machine translation. *Applied and Computational Engineering*, *37*(1), 95–101. https://doi:10.54254/2755-2721/37/20230479

[57] Way, A. (2018). Quality Expectations of Machine Translation. *Translation Quality Assessment*, 159–178. https://doi.org/10.1007/978-3-319-91241-7_8

[58] Weaver, W. (1955). Translation. *Mach Transl Lang*, *14*, pp. 15-23.

[59] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa. H., ..., Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.* https://doi.org/10.48550/arXiv.1609.08144

[60] Zaghlool, Z. D. M., & Khasawneh, M. A. S. (2023). Aligning Translation Curricula with Technological Advancements; Insights from Artificial Intelligence Researchers and Language Educators. *Studies in Media and Communication*, *12*(1), 58. https://doi.org/10.11114/smc.v12i1.6378

Mozhgan Ghassemiazghandi holds a PhD in Translation and currently serves as a Senior Lecturer at the School of Languages, Literacies, and Translation at Universiti Sains Malaysia. Her areas of interest include Audiovisual Translation, Machine Translation, and Translation Technology. Mozhgan is also a translator and subtitler with over a decade of experience.