Six NMT Systems, One Language Pair: Which Best Translates Arabic-English?

Rand Habib[1], Linda Alkhawaja[1], Ogareet Khoury[1], Sa'ida Al-Sayyed[1]

[1] English Language Department, Al-Ahliyya Amman University, Jordan

Correspondence: Linda Alkhawaja, English Language Department, Al-Ahliyya Amman University, Jordan. E-mail: l.alkhawaja@ammanu.edu.jo

Received: January 24, 2025 Accepted: May 20, 2025 Online Published: July 11, 2025

doi:10.5430/wjel.v16n1p1 URL: https://doi.org/10.5430/wjel.v16n1p1

Abstract

This study evaluates the quality of translations produced by six different Neural Machine Translation (NMT) systems when translating from Arabic to English. The systems under study are Google Translate, Microsoft Bing, Yandex, Systran, ChatGPT-4, and Amazon Translate. Given the precision and complexity of the Arabic language, the study aims to examine the most effective NMT system and understand how translators can utilize these tools. To achieve the study's objectives, 1,000 Arabic sentences and their English translations are examined, with established translations verified by human translators used as reference benchmarks for evaluating machine translations. Data are collected from the Tatoeba platform (2024), accessible to researchers online, and are analyzed using the Bilingual Evaluation Understudy BLEU system. The study's findings reveal significant variations in translation quality among the systems tested, highlighting the necessity for translators to be involved in the machine translation editing process. Moreover, the results indicate that ChatGPT-4 outperform other systems in producing high-quality translations. This study contributes to translation studies by offering a comprehensive comparative analysis of current NMT systems, providing practical insights for translators, and advancing research on machine translation applications.

Keywords: machine translation, quality assessment, bleu metric, assessment competency, artificial intelligence

1. Introduction

Neural machine translation (NMT) is an approach to machine translation that uses an artificial neural network to imitate the structure and functioning of the human brain (Kenny, 2022). The NMT system proved to be more reliable than the older Statistical Machine Translation (SMT) model, with a 60% reduction in translation errors (Kenny, 2022). This is because NMT was designed to copy the human brain's neural network in order to create links between common words and phrases. Translation quality has drastically improved with the use of Artificial Intelligence (AI). However, different languages benefit from this technology in varying degrees (Alissa, 2024). Languages with intricate structure are harder to be translated correctly using machine translation (Maučec and Donaj, 2019; Alkhawaja, 2023). An example would be translating texts from English to Arabic. This is due to the intricate rules and structure of the Arabic language when compared to the morphologically simpler structure of the English language. Even with years of improvement to this technology, there are still challenges to be solved within the field of machine translation, including addressing uncommon words, developing translations for less used languages, and ensuring that the translated text remains coherent and flows (Maučec & Donaj, 2019). The study aims to build on earlier research in the fields of translation studies and artificial intelligence. While many studies have examined the output of various MT software for the English-Arabic language pair-including Google Translate, Systran, Babylon, The Translator, Sakhr, Al-Mutarjim TM Al-Araby, and Systran (e.g. Almahasees, 2017; Jabak, 2019)—to the best of the researchers' knowledge, no one has conducted such a large-scale comparison. To accomplish this comparison, a corpus of 1000 English sentences and their corresponding Arabic translations were examined, sourced from the open online platform Tatoeba. Additionally, the data were analyzed electronically using the BLEU metric system to measure the accuracy and fluency of the translations. The various limitations of NMT systems in translating from Arabic to English warrant the need for a critical investigation of this type of analysis. The main purpose of the study is to conduct a thorough investigation of various NMT systems to assess the quality and performance of each system. The findings of this study are expected to provide useful information to translators to enable them identifying the best machine translation system to use in the workplace. On that basis, this study seeks to answer the following research questions; to what extent do human translators benefit from using NMT in their work? And which system among Google Translate, Microsoft Bing, Yandex, Systran, ChatGPT-4, and Amazon Translate is the most effective in translating general sentences from Arabic to English?

2. Literature Review

Various research studies (Almahasees, 2017; Jabak, 2019; Popel et. al. 2020; Zakraoui et al. 2021; Beseiso et. al. 2022), have delved into the effectiveness of machine translation applications and software, in areas like literature, media, news, politics and business. For example, Almahasees (2017) compared two MT translation systems; Google Translate and Microsoft Bing to determine which one excels in producing Arabic-to-English translations. The study adopted the automatic evaluation metric BLEU to evaluate the translations of 25 sentences in Arabic sourced from Jordan's Petra News Agency. Also, the English translations of these sentences were taken from Jordan's Petra News Agency and used as a benchmark. The findings revealed that Google Translate performed better than Microsoft Bing when translating Arabic political texts into English.

Similarly, Jabak (2019) conducted a study to assess the quality of Google Translate for Arabic to English translations. The study identified lexical and syntactic errors in the translations, with lexical errors being more prevalent than syntactic errors. The study concluded that employing human post-editing is indispensable when using Google Translate for Arabic to English translation.

Another study conducted by Alkhawaja et al. (2024) examined the quality of Google Translate and ChatGPT 3.5 by analyzing 1000 sentences sourced from Tatoeba database. The translations were evaluated by three bilingual experts in linguistics and translation using a quality assessment (QA) model. The results revealed that translation inaccuracies were the most common errors, followed by sentence distortions and orthographic mistakes. The study concluded that ChatGPT outperformed Google Translate in Arabic-English machine translation, while also noting that the translations provided by Google Translate were acceptable in terms of adequacy and accuracy.

Abdelaal and Alazzawie (2020) examined translation errors and assessed the fluency and semantic adequacy of Google Translate output. Moreover, they evaluated the extent of effort needed by a human translator to rectify the translation. Data from online newspapers were analyzed using a mixed-method approach, employing Hsu's classification of machine translation errors, Multidimensional Quality Metrics, and Localization Quality Evaluation. The study evaluated semantic adequacy and fluency using a questionnaire adopted by Dorr, Snover, and Madnani (2011). The findings indicated that lexical errors, such as omission, semantic errors, and inappropriate lexical choices, are the most common errors. Based on this, it can be concluded that while Google Translate is useful for translating Arabic informative news texts, it is not fully effective and still falls short of achieving the quality level of human translators.

In the same vein, Almahasees (2021) assessed the performance of three MT applications—Microsoft Bing, Google Translate, and Sakhr—across different linguistic aspects and text genres. The study concluded that Google Translate achieved the highest scores in terms of accuracy and fluency when translating texts from organizations such as the UN, UNESCO, and WHO, while Microsoft Bing excelled in translating from English to Arabic. Lastly, Sakhr was particularly effective in translating single words, sentences, and idioms.

Moreover, a study conducted by Sismat (2022) examined the quality of two machine translation systems, Google Translate and Microsoft Bing. The study revealed error patterns in the translation of four Arabic texts, with Google Translate and Microsoft Bing scoring 72.2% and 73.1%, respectively. The study identified specific errors, including meaning inaccuracies, syntactic mistakes, and terminology issues. It concluded that Google Translate demonstrated slightly better accuracy compared to Microsoft Bing.

Khondaker, Waheed, Nagoudi & Abdul-Mageed (2023) also conducted a study on the efficiency of ChatGPT 3.5 in translating texts from/into Arabic and English languages. They examined a corpus of texts from various genres such as literary, historical, journalistic, scientific and legal texts. They found that ChatGPT3.5 produced high quality translations across these genres; however, the translations need the intervention of human translators.

On that basis, one could argue that ChatGPT can be considered as a useful machine translation for various genres. Many other studies have been conducted on different language pairs and various text types and genres to examine the quality of translation produced by various MT systems, applications, and translation tools (e.g., Sanz-Valdivieso & López-Arroyo, 2023; Ameen & Ahmed, 2023; Shalabi & Amrieh, 2024). Based on the previous literature review, it is evident that MT is an effective tool for translators. However, it still exhibits several types of errors, and its quality varies between systems, with some performing better than others. To the best of the researchers' knowledge, none of the previous studies have conducted a comprehensive comparison among a relatively large number of NTM systems in comparison to human translation. Table 1 summarizes the scope of previous studies, showing that most comparisons are limited to a maximum of three MT tools, whereas the current study compares six tools, providing a more comprehensive analysis.

Table 1. Comparison of machine translation tools in previous studies

Name of Study	Names of Programs	Number of Programs Compared
Almahasees (2017)	Google Translate and Microsoft Bing	2
Jabak (2019)	Google Translate	1
Alkhawaja et al. (2024)	Google Translate and ChatGPT 3.5	2
Abdelaal and Alazzawie (2020)	Google Translate	1
Almahasees (2021)	Microsoft Bing, Google Translate, Sakhr	3
Sismat (2022)	Google Translate and Microsoft Bing	3
Khondaker et. al. (2023)	ChatGPT 3.5	1

This study makes a significant contribution to the field of machine translation (MT) by conducting a comprehensive comparison of six widely used MT tools, a notable advancement over previous studies that typically limit their analysis to a maximum of three tools (as illustrated in Table 1). By expanding the scope of comparison, this research provides a more holistic evaluation of MT performance, particularly in translating between Arabic and English. Moreover, the findings offer valuable guidance to translation professionals, allowing them to choose the most suitable tools for different translation tasks. This study not only fills a gap in existing literature but also sets a new benchmark for future research in the evaluation of MT tools, especially for a linguistically rich and structurally complex language like Arabic (Shalabi & Abu Amrieh, 2025; Abu Manie et al. 2025).

3. Theoretical Framework

3.1 Neural Machine Translation (NMT)

In our increasingly interconnected world, NMT is at the forefront of language technology, revolutionizing the way we overcome linguistic barriers (Xiao et al., 2023). NMT models follow the common sequence-to-sequence learning architecture (Chen., 2022). It consists of an

encoder and a decoder Recurrent Neural Network (RNN) which the encoder transforms the input sentence into a list of trajectories, one vector per input, and the decoder produces one output at a time until the special end-of-sentence token is produced (Fernandes et al., 2022). The NMT process works by preprocessing, whereby text data were converted to lowercase, transformed into numerical representations, and tokenized (Yin, Li, Meng, Zhou, & Zhang, 2023). A trained NMT model can be able to translate new sentences by decoding the translation word by word, encoding the input sentence into context vectors and using context vectors for generating words to predict the subsequent word (Fernandes et al., 2022). Training NMT model involves a process that includes two main steps: gathering broad parallel corpora (pairs of sentences in the source and target language) and utilizing these corpora to train the encoder-decoder network using direct learning. Training the model to minimize the differences between its expected translations and the actual translation is done by applying the optimization techniques such as gradient descent and back propagation. Systemization methods and dropout are used to moderate overfitting (Yin et al., 2023). In addition, the output sequence is subsequently post-processed revert tokenization, applying necessary capitalization and punctuation to combine sub-words into complete words (Fernandes et al., 2022). NMT is a valuable source for professionals from different fields such as business, healthcare, politics, and education. For example, in the healthcare field, many doctors and nurses use NMT to communicate with their patients who may not speak the same language. This can enhance the medical field or sector (Kl ímov áet al., 2022). In conclusion, NMT offers different stages of accuracy that can be considered as a turning point in the development of machine translation. This study will focus on the quality of NMT output, testing six of the most well-known systems, by evaluating translations from Arabic to English.

3.2 Translation Quality Assessment

Translation Quality Assessment (TQA) is a fast-growing field and is defined to measure the quality of translation (Putri, Sofyan, & Nasution 2022). A good translation should not add different ideas, include immaculate information, or confuse the reader (Cui, Liu, & Cheng, 2023). Translation quality assessment can be divided into two categories: qualitative and quantitative assessments. Qualitative assessment measures the quality of translations using established indicators but does not provide an exact score, while quantitative assessment presents the assessment results in numerical form (Cui et. al, 2023). The 'scoring method' is one way to quantitatively evaluate translations. This is where an evaluator assigns a score based on a reference translation or an ideal translation according to the evaluator's perspective. The final score for the entire translation is the sum of the scores for different parts as per the translation scoring model (Wang & Daghigh, 2023).

The 'statistical method' is another commonly used quantitative way to assess machine translation. This method requires the evaluator to find manually or semi-automatically equivalence between the reference translation and original text. The evaluator scores the results of the mutually matching degrees whereby an increase in matches results in an increase in translation quality. This method only works if there are available high-quality reference translations available. Both the 'scoring method' and the 'statistical method' require reference translations as scoring benchmarks (Abanomey & Almossa, 2023). Computer tools are currently being used to evaluate the quality of the translation and also to identify any major contradictions during the actual quality assessment process. This is done to achieve three main purposes such as to verify correct spelling and punctuation, guarantee that the appropriate terminology has been used and to detect any formatting errors and repeated words or multiple spaces.

The goal of all translation quality assessments is to identify a method that can impartially and effectively evaluate the quality of different types of translations. This includes the process of assessing the consistency and accuracy of translations compared to the original text at both macro and micro levels, as well as providing feedback on the strengths and weaknesses of the translated material (Wang & Gu, 2024). The assessment process must also focus on comparing the original text with the translated version, and the assessment findings should be based on specific standards and criteria. Therefore, the increase of impartial, precise, and practical assessment standards and criteria is crucial for creating a reliable translation quality assessment model (Wang & Gu, 2024).

3.2.1 BLEU Metric

BLEU (Bilingual Evaluation Understudy) is used to assess the quality of machine-translated texts and compare it to human-translated texts. It can measure the similarity between original texts and their machine translations using n-grams (Datta, Joshi & Gupta, 2022). An n-gram symbolizes a set of consecutive words in a sentence (Ashraf, 2023). The concept of n-gram is globally used in standard text processing and is not only limited to the field of BLEU metric. n-grams are serial groups of words such as bigrams, unigrams and trigrams (Ashraf, 2023). They measure the actual precision of translations by simply conducting a comparison with reference translations, and by calculating the degree of similarity between them. Higher BLEU scores indicate greater similarity between the two texts, thus, higher quality translations.

For example, in the phrase "This is a big data AI book", the n-grams are:

- •1-gram (unigram): "This," "is" "big," "data " "AI" "book".
- •2-gram (bigram): "This is" "is big" "big data " "Data AI" "AI book"
- •3-gram (trigram): "The is big" "is big data" "big data AI" and "Data AI book"
- •4-gram: "This is a big data AI book"

The order of words in an n-gram is crucial, so "This is a big data AI book " would not form a valid 4-gram. N-grams in the BLEU metric measure different aspects in translation. To illustrate, 1-gram measures the accuracy of individual word translations, 2-gram measures the accuracy of pairs of consecutive words and 3-gram measures the accuracy of triplets of consecutive words. Finally, 4-gram measures the accuracy of quadruplets of consecutive words (Datta et al., 2022).

This is the formula for a BLEU score:

$$BLEU = BP * exp (\sum pn)$$

BP is a short-term for brevity penalty, which acts as a major factor in scoring translations that fall short of the length of the reference text. BP is also used to adjust its score and make it more accurate in situations where the result shows that the translation is shorter than the reference text (Ni'mah, Fang, Menkovski, & Pechenizkiy., 2023). The penalty is determined by (reference length / translated length), with reference length representing the total word count in reference text and translated length representing the total word count in the machine-generated translation. Moreover, the precision of n-grams (pn) quantifies the precision of n-grams by evaluating the ratio of shared n-grams between the machine-generated translation and the reference text, compared to the total count of n-grams in the machine-generated translation (Ni'mah et al., 2023).

3.3 Machine Translation

Machine translation is an excellent starting point for any translation project, mostly for large volumes of text or texts written in well-ordered sublanguages, such as weather forecasts (Dalibor, 2024). Moreover, MT is essential for any business looking to expand beyond its local market. Companies must consider how to market their products in other countries or regions, often requiring translation into multiple official languages to serve diverse communities. However, MT is also considered to be effective for texts that do not require precision or creativity, such as user-generated content and technical documentation. Which can be automatically translated and then post-edited by human translators for accuracy and consistency. This allows you to deliver projects more quickly and of higher quality at a fraction of the cost of manually translating the project (Chen, 2022). This research focuses on assessing the translation quality of specific studied software systems: Google Translate, Microsoft Bing, Yandex, Systran, ChatGPT 4, and Amazon Translate.

3.3.1 Google Translate

Google Translate is a well-known translation platform which provides all kinds of on spot text translation services using multiple technologies such as NMT (Sanz Valdivieso & Arroyo, 2023). By using its unique type of algorithms and user-friendly interface, Google Translate has quickly become a major tool for business, academics, and individuals. With its wide language support and continuous improvements, it helps to promote cross-cultural exchange and bridge language gaps (Shukla, Bansal, Badhe, Ranjan, & Chandra, 2023).

Google Translate has impinged communication within English and Arabic-speaking communities, facilitating interactions among individuals, government entities, and various organizations. It facilitated interactions and connections among people from diverse cultural backgrounds, proving irreplaceable for business negotiations, educational purposes, or casual conversations (Al-Sabbagh, 2023).

3.3.2 Microsoft Bing

The search mechanism known as Microsoft Bing doubles as a robust translating device besides its prime function of fetching data (Mart ń, 2017). Its proficiency in converting English into Arabic is hailed for breaking down the barriers to communication that exist across different cultures (Almahasees, 2017). Ever since it debuted in the year 2009, there has been continuous refinement and incorporation of avant-garde technology by Microsoft Bing, all aimed at elevating its efficiency and making the user experience more fulfilling (Mart ń, 2017).

To improve the precision and overall quality of translations, Bing Translator has adopted NMT technology. This advancement proves particularly advantageous when dealing with language duos where subtleties are crucial for effective communication (Rescigno, Monti, Way, & Vanmassenhove, 2020).

Microsoft Bing Translator offers a range of features beneficial to individuals from various backgrounds. The NMT algorithms deliver translations that flow naturally within their contexts. It serves the needs of people across sectors worldwide, helping to bridge language barriers effectively (Sismat, 2022). Whether it is articles, business papers, or casual chats on platforms, this tool facilitates communication among individuals speaking diverse languages. This promotes collaboration across continents and nurtures a sense of unity by fostering an understanding of different perspectives (Ahmed & Lenchuk,2024).

The English to Arabic translation feature on Microsoft Bing represents a significant advancement in translation technology Nevertheless, continuous research and progress are important to enhance translation standard and pursue the challenges posed by cultural nuances and linguistic intricacies.

3.3.3 Yandex

Yandex is a web service that offers a range of services, primarily translation, research functions, and maps. Yandex can effectively overcome language barriers (Lengkong, Mandias, & Tombeng, 2022). It has made significant advancements in handling language pairs where cultural differences are pronounced, gradually improving the accuracy and nuances of its translations by integrating NMT technology and extensive datasets into its algorithms (Adawiyah, Baharuddin, Wardana, & Farmasari, 2023). Moreover, Yandex facilitates communication between English and Arabic speakers, extending beyond simple chat translations. Numerous studies suggest that since its introduction, Yandex has utilized machine learning technologies and advanced language processing to enhance the quality and fluency of its translations (Kalinina & Kalinina, 2023).

3.3.4 Systran

Systran is an automated translation system that can bring intercultural communication and overcome linguistic challenges that appear in the

translation process (Musaad & Towity, 2023). It also offers various linguistic solutions for translators, producing translations from/into English and Arabic that are both accurate and reliable (Musaad & Towity, 2023). Systran includes various features to meet the needs of translators and language service providers such as fast and accessible translations of webpages (De Oliveira & Anastasiou, 2011).

Systran provides translations for the Arabic-English language pair, making it easier for Arabic and English-speaking communities to communicate. It offers translations for manuals, business documents, health, and educational texts, enabling users to overcome language barriers (Ismailia, 2023). According to Zughoul and Abu-Alshaar (2005, p. 3), Systran is "the leading provider of the world's most scalable and segmental translation architecture". It offers revolutionary translation solutions for the Internet, PCs, and network infrastructures, facilitating communication in 36 language pairs across 20 specialized domains.

3.3.5 ChatGPT-4

ChatGPT is an AI chatbot developed by OpenAI, powered by advanced large language models, and has garnered significant interest across various industries (Ghosh & Caliskan, 2023). In March 2023, OpenAI announced the release of GPT-4, the highly anticipated fourth generation of their GPT language model series. According to OpenAI, this new model includes visual input capabilities and enhanced natural language abilities (Teebagy, Colwell, Wood, Yaghy, & Faustina 2023). ChatGPT-4 is designed to handle complex translations, delivering impressive performance when translating diverse and longer texts (Gala & Makaryus, 2023). Additionally, ChatGPT-4 is not only built for the translation industry but also functions through prompts, making it an effective tool for translators (Kadaoui et al., 2023). Moreover, ChatGPT-4 excels at translating words, idioms, and sentences, considering the historical and cultural contexts of texts, as well as maintaining flawless linguistic coherence (Jiao, Wang, Huang, Wang, & Tu, 2023).

ChatGPT-4 can handle technical errors, terms and translate complex documents to come up with a fathomable language that is useful for translators working with texts in specialized fields, such as legal, medical or media documents. As for these documents, ChatGPT-4 can also provide translators with a plain-language explanation for a sentence, paragraph or a full text took from a legal document. This helps the translators to ensure source text (ST) translation has an accurate meaning in the target text (TT) (Peng et al., 2023). For these reasons, ChatGPT-4 helps translators not only to understand the text but also conduct research and familiarize themselves with specific terminologies before embarking on diverse translation tasks. One possibility is prompting the model to propose alternative word choices for the translator's own work or machine-generated translations, using it as a supplement to collocation dictionaries, with context awareness. ChatGPT can provide various suggestions to refine translations, improving their accuracy, clarity, and fluency (Sallam & Mousa 2024). Therefore, ChatGPT and similar AI tools can improve the work of translators and language professionals. These systems may function as successful learners rather than just tools, augmenting human intelligence and amplifying professional skills. By adopting a collaborative approach that organizes work and defines roles, translators can leverage AI to increase productivity and creativity while maintaining control over the translation process.

3.3.6 Amazon Translate

Amazon Translate leverages various technologies, including deep learning, to process input text and generate accurate translations. It employs a neural network to analyze the context of the source text, considering how words in a sequence influence one another (Zhivotova, Berdonosov & Redkolis, 2020). This approach produces more precise translations compared to statistical and rule-based models, which only consider the context of a few words at a time. Consequently, Amazon Translate can better convey the original meaning and context of the source text (Moneus & Sahari, 2024). Amazon translate is a neural machine translation service that uses cutting-edge machine learning technologies to deliver high-quality translations (Callison-Burch, 2009). This machine translation has proven to be beneficial in sectors where accurate translation is essential, such as e-commerce customer assistance and global diplomacy (Gaurav et al., 2013). Amazon Translate supports up to 75 languages and offers 5,550 translation combinations, showcasing its advanced capabilities (Keary, 2024). One example of a supported language combination is Arabic to English. By using such translation tools, Arabic speakers can easily participate in activities in English, such as online shopping, business transactions, and customer service interactions (Gaurav et al., 2013).

4. Materials and Methods

This study adopts a quantitative research methodology to conduct a comprehensive analysis to evaluate translation precision for the NMT systems. The analysis includes the evaluation of six renowned NMT applications: Microsoft Bing, Yandex, Systran, ChatGPT 4, and Amazon Translate, in translating from Arabic to English.

4.1 Software Selection

Many strategic reasons were considered while choosing these six prominent applications. Firstly, each selected application is effective in translating the Arabic-English language pair, considering the Arabic complexity according to many studies. Secondly, these six systems use a neural network model to translate from one language into another. Yandex and Systran are chosen for their impressive approach in training their own machine translation models using their own data which can increase translation accuracy for specific industries (Vanjani & Aiken, 2020). Similarly, Google Translate and Microsoft Bing are well-known machine translations with a huge user base and inclusive language support, which make them indispensable for comparison (Alkatheery, 2023). ChatGPT 4 uses the approach of natural language production and perception and, thus, distinguishes its translation output. Amazon Translate has a high-performance and scalable infrastructure that helps to evaluate immersion in real performance (Diab, 2021). On that basis, the selected tools are expected to provide a comprehensive framework for assessing the quality of NMT output when translating texts from Arabic to English.

4.2 Data Collection Process

Due to constraints posed by the limited availability of Arabic datasets and computing resources, our data collection focused on a subset of 1,000 randomly selected sentences. These sentences were extracted from Tatoeba (2024), a large database of sentences and translations. The collected source sentences will be translated into English using six machine translation systems. Both the original sentences and their translations will be evaluated using the BLEU metric. The quality assessment scores generated by BLEU will then be compared and discussed. Finally, the performance of the six-machine translation systems will be juxtaposed with that of human translation.

4.3 Data Analysis

There are various evaluation systems to assess the quality of MT with the BLEU metric being the most common (Warner, 2022). According to (Marie et. al. 2021, p. 7299) "the overwhelming majority of MT publications uses BLEU. Precisely, 98.8% of the annotated papers report on BLEU scores." Moreover, Rivera-Trigueros (2022) asserts that BLEU is the most common metric for MT evaluation. It aims to compare the quality of different MT systems by evaluating their output against human translation as a reference. Human reference translation is a fundamental requirement to employ in the BLEU metric (Rossi and Carré, 2022). The BLEU metric is designed to evaluate translation quality in terms of its adequacy and fluency. It measures the translation's lexical precision by calculating word level matches. The BLEU n-grams, f-measure, recall, and precision are used to assess translation quality with higher values indicating better translation quality. That is, the n-grams score measures how close the output from an MT system is to a professional human translation of the same text. BLEU largely uses the modified n-gram precision approach to indicate whether the outputs of MT are good or not (Warner, 2022). This is calculated by taking the number of n-grams in the translated text being tested and dividing it by those that are found to match this specific translation and its source. The precision of each n-gram order is measured, and these precisions are then geometrically averaged to obtain the final number. BLEU normally employs the customary maximum n-gram order, which is a string of four words. The measure calculates a modified precision rate that has been adjusted with a brevity penalty to discourage short sentence usage rather than the reference sentence (Rossi and Carré, 2022).Although there are alternative metrics like METEOR, TER, BLEU, and NIST, the consensus among machine translation experts is that BLEU is the most frequently used and accurate metric (Maučec and Donaj, 2019; Warner, 2022). BLEU provides us with an objective and consistent means of assessment and allows us to compare the quality of translations generated by each system. Specifically, we will be using BLEU scores to compare the accuracy of translations from Arabic to English. These scores will be the foundation of this comparative evaluation. The data was analyzed using the Interactive BLEU tool on the Tilde.ai website.

5. Results

The findings of this study are expected to assist translators, whether freelancers or those working for language service providers, in choosing the best systems for their work, based on results from the BLEU metric n-grams. The findings indicate varying levels of effectiveness among all systems, as reflected in their BLEU scores. ChatGPT achieved the highest BLEU score, 59.11, indicating superior accuracy and fluency. Google Translate also performed well, with a BLEU score of 56.76, demonstrating its ability to capture nuances and produce contextually appropriate translations. Similarly, Microsoft Bing achieved a BLEU score of 52.9, indicating effectiveness in translation, though slightly less than ChatGPT and Google Translate. Amazon Translate's BLEU score of 54.18 indicated a very good performance, particularly in handling large volumes of text efficiently. Moreover, Yandex BLEU score is 53.59; a score that is lower than the previous systems. Finally, Systran's BLEU score of 51.71 was the lowest among the evaluated systems. Therefore, BLEU scores for the six NMT systems, showing that ChatGPT has the highest overall BLEU score of 59.11, indicating it generally provides the most accurate translations. Google Translate follows with a score of 56.76, also demonstrating high translation quality. Amazon and Yandex have similar overall scores of 54.18 and 53.59, respectively. However, the results show that Bing and Systran have lower overall scores of 52.90 and 51.71, respectively. Based on BLEU score interpretation (see figure 1), All the above systems achieved a score between 51 and 59, indicating very high-quality translations. These findings support the Alternative Hypothesis (H1) that human translators benefit significantly from using NMT, as the high-quality translations can assist translators in their work.

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

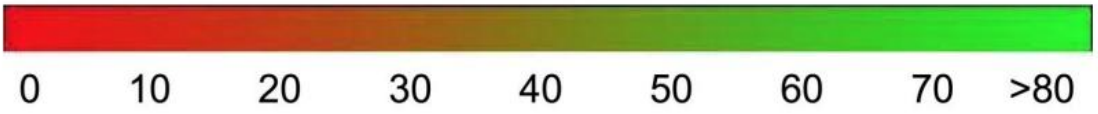


Figure 1. Interpretation of the BLEU Score (Evaluating Models | AutoML Translation Documentation | Google Cloud)

5.1 ChatGPT and Google Translate

Based on the BLEU metric scores, the results of comparing ChatGPT and Google Translate with human translation are presented in Figures 2 and 3. ChatGPT 4, represented in blue, has a higher overall BLEU score of 59.11 compared to Google Translate with a score of 56.76, shown in green.

Across all n-gram categories (1-gram, 2-gram, 3-gram, and 4-gram), figure 2 presents that ChatGPT-4 consistently outperforms Google Translate indicating superior translation accuracy and fluency. ChatGPT-4 achieves 83.21 for 1-gram, 65.29 for 2-gram, 53.25 for 3-gram, and 44.35 for 4-gram. In contrast, Google Translate scores 82.56 for 1-gram, 63.80 for 2-gram, 51.20 for 3-gram, and 41.64 for 4-gram.

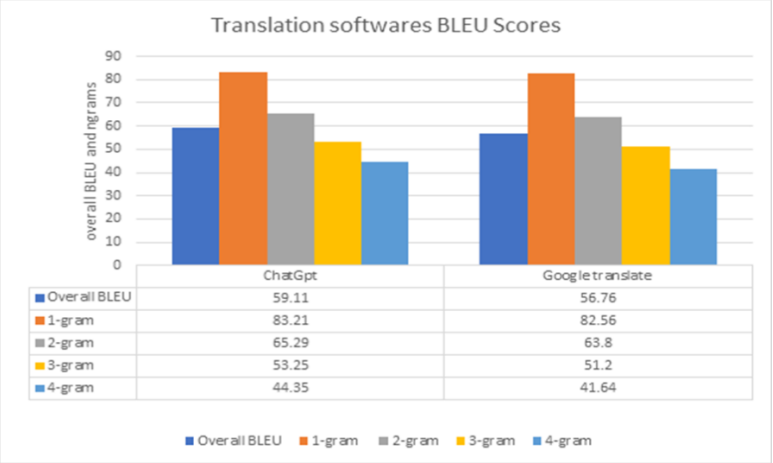


Figure 2. BLEU scores for ChatGPT and Google Translate- n-grams based

Figure 3 highlights that ChatGPT-4 frequently achieves higher individual sentence BLEU scores, reinforcing its overall better performance in providing accurate and contextually appropriate translations.

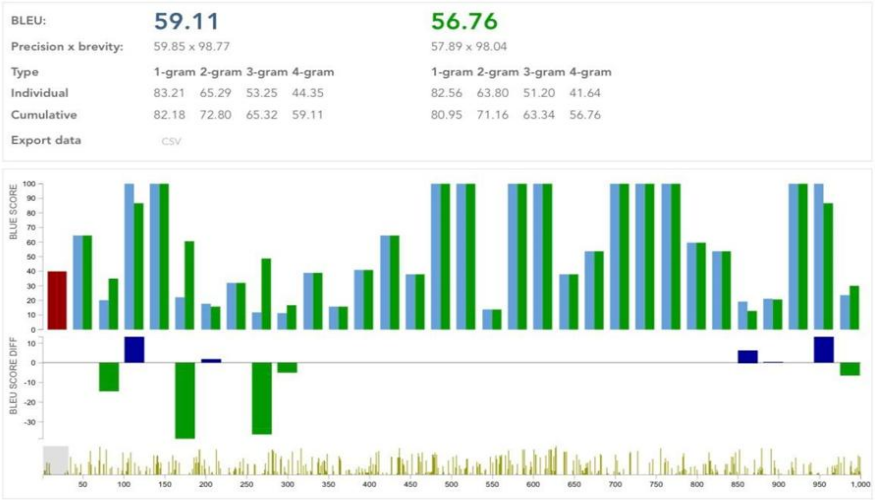


Figure 3. BLEU scores for ChatGPT and Google Translate- n-grams based

To illustrate the observed variations in translation quality between ChatGPT-4 and Google Translate, the following examples are taken randomly from the BLEU metric system.

Table 2. Example (1) ChatGPT-4 vs Google Translate

Sentence 28	BLEU Score	ST
Human	100	Cancer can be cured easily if it is found in its first phase.
ChatGPT-4	18.99	Cancer can be treated if discovered in the early stages.
Google Translate	12.57	You can treat cancer if it is discovered in the early stages.

Table 3. Example (2) ChatGPT-4 vs Google Translate

Sentence 447	BLEU Score	ST
Human	100	How easily one acquires bad habits!
ChatGPT-4	16.59	How easy it is for one to acquire bad habits!
Google Translate	14.45	How easy it is for a person to acquire bad habits!

Table 4. Example (3) ChatGPT-4 vs Google Translate

Sentence 896	BLEU Score	ST
Human	100	Some are moderate; some are radical.
ChatGPT-4	18.52	Some of them are moderate, and others are radical.
Google Translate	11.73	Some of them are moderate, others are extremist.

5.2 Amazon Translate and Yandex

The results of comparing Amazon Translate and Yandex with human translation are presented in Figures 4 and 5. The findings reveal that the BLEU score for Amazon Translate, represented in blue, is slightly higher at 54.18 compared to Yandex, which is shown in green, at 53.59.

Regarding the n-gram scores presented in figure 4, Amazon Translate scores 79.77 for 1-gram, 60.25 for 2-gram, 47.33 for 3-gram, and 37.87 for 4-gram. In contrast, Yandex scores 79.45 for 1-gram, 59.85 for 2-gram, 47.14 for 3-gram, and 37.88 for 4-gram. While both tools show nearly similar performance, Amazon Translate has slightly higher overall scores in almost all n-grams, indicating marginally better translation accuracy and fluency.

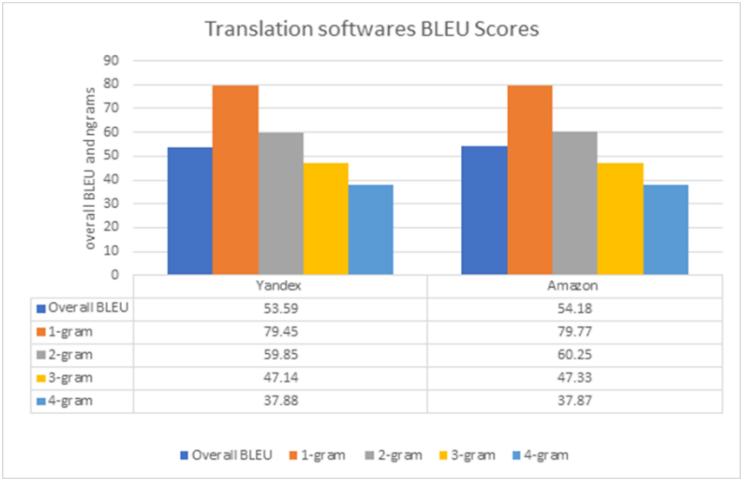


Figure 4. BLEU scores for Yandex and Amazon - n-grams based

Figure 5 further highlights that Amazon Translate frequently achieves higher individual sentence BLEU scores, reinforcing its better overall performance in providing accurate and contextually appropriate translations.

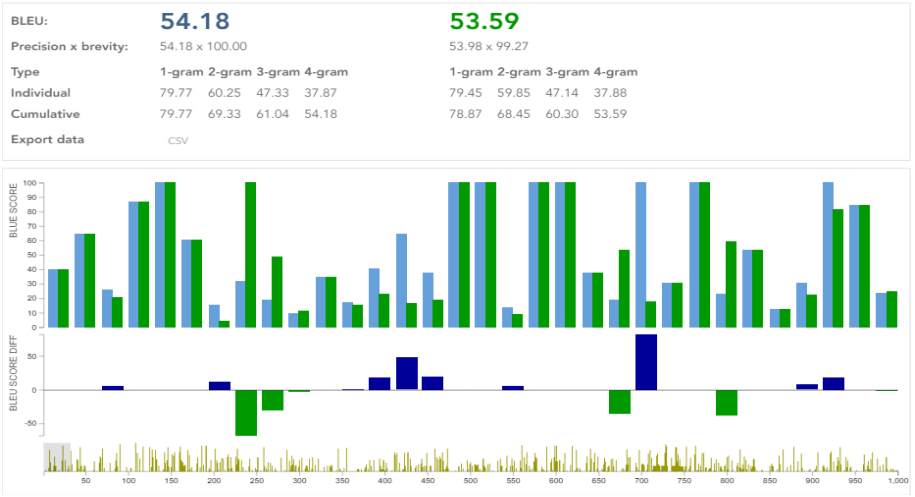


Figure 5. BLEU scores for Amazon Translate and Yandex – sentence-based

To illustrate the observed variations in translation quality between Amazon Translate and Yandex, the following examples are taken randomly from the BLEU metric system.

Table 5. Example (1) Amazon Translate vs Yandex

Sentence 103	BLEU Score	ST
Human	100	I believe Shakespeare is the greatest dramatist that ever lived.
Amazon Translate	26.20	I think Shakespeare is the greatest drama writer in history.
Yandex	38.09	I think Shakespeare is the greatest dramatist in history.

Table 6. Example (2) Amazon Translate vs Yandex

Sentence 279	BLEU Score	ST
Human	100	It is no wonder that some people feel anxiety at the thought of walking into a hospital.
Amazon Translate	47.15	It is no wonder that some people feel nervous when they are admitted to the hospital.
Yandex	4.76	no wonder some feel nervous when they are hospitalized.

Table 7. Example (3) Amazon Translate vs Yandex

Sentence 537	BLEU Score	ST
Human	100	The temperature ranges from thirty to forty degrees Celsius.
Amazon Translate	22.037	The temperature in summer is thirty to forty degrees.
Yandex	40.44	The temperature in summer ranges from thirty to forty degrees.

5.3 Microsoft Bing and Systran

Based on the results obtained from the BLEU metric system, Figures 6 and 7 show that Microsoft Bing, represented in green, has a higher overall BLEU score of 52.90 compared to Systran, which has a score of 51.71, shown in blue.

For the n-gram scores, Microsoft Bing scores 80.45 for 1-gram, 60.08 for 2-gram, 47.09 for 3-gram, and 37.92 for 4-gram. Systran scores 80.88 for 1-gram, 59.74 for 2-gram, 46.52 for 3-gram, and 37.30 for 4-gram. While Systran has a higher score in the 1-gram category, Microsoft Bing scores higher in the other n-gram, thus, in the overall quality of n-grams and translations.

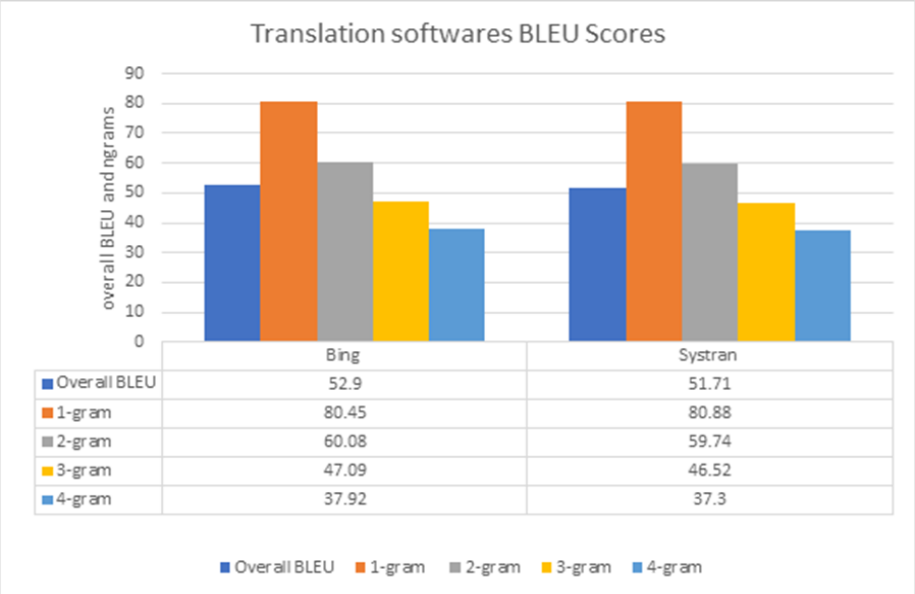


Figure 6. BLEU scores for Microsoft Bing and Systran- n-grams based

Figure 7 highlights that ChatGPT-4 frequently achieves higher individual sentence BLEU scores, reinforcing its overall better performance in providing accurate and contextually appropriate translations.

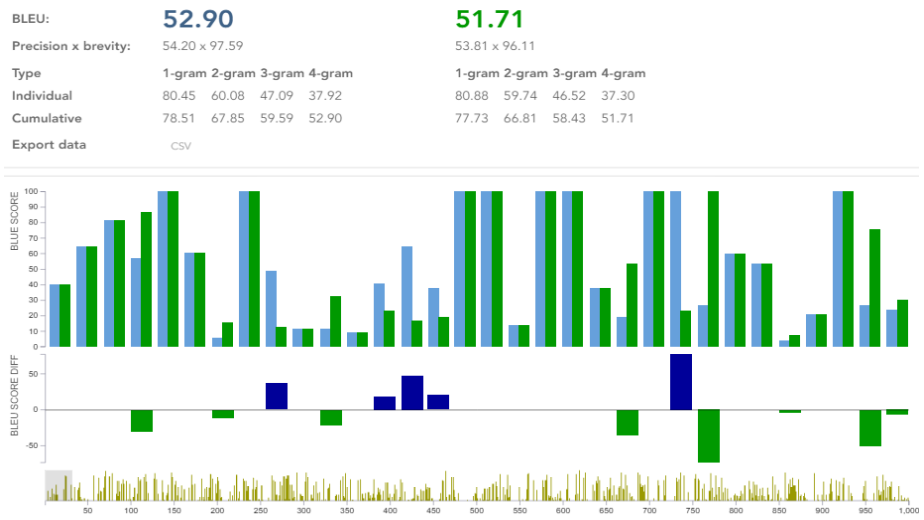


Figure 7. BLEU scores for Microsoft Bing and Systran- sentence-based

Tables 8, 9 and 10 present examples that are taken randomly from the system. They demonstrate the variations in translation quality between Microsoft Bing and Systran.

Table 8. Example (1) Microsoft Bing vs Systran

Sentence 391	BLEU Score	ST
Human	100	Young men and women fought to defend their country.
Microsoft Bing	61.56	men and women fought to defend their homeland
Systran	60.34	Men and women who fought to defend their country.

Table 9.1 Example (3) Microsoft Bing vs Systran

Sentence 474	BLEU Score	ST
Human	100	Suddenly rain began to fall.
Microsoft Bing	59.46	Suddenly the rain began to fall.
Systran	27.05	Suddenly the rain started to fall.

Table 10. Example (2) Microsoft Bing vs Systran

Sentence 514	BLEU Score	ST
Human	100	Tom's parents seemed relieved to hear that he was safe.
Microsoft Bing	17.67	Tom's parents were reassured when they heard that he was okay.
Systran	14.45	Tom's parents were reassured when they heard he was fine.

6. Discussion

The results revealed that human translators can greatly benefit from NMT in their work. These advantages manifest in various ways, most importantly in enhancing efficiency and speed. While human translators need to dedicate considerable time to creating translations, NMT systems can quickly produce translations shortly after receiving the original texts. This allows translators to focus more on refining and editing, thereby enhancing overall productivity. Moreover, NMT can ensure uniformity in terminology and phrasing, which is especially valuable for maintaining consistency in technical documents, legal texts, and other specialized content.

NMT is also very helpful in managing repetitive or highly structured texts and effectively handling bulk translations with high proficiency. Human translators can then review and edit these translations to ensure accuracy. This process often proves to be faster than translating everything from scratch. Many translators find NMT to be a valuable resource for initial drafts, especially in contexts where speed is crucial, such as time-sensitive reports.

Due to the fact this research focuses solely on the language pair Arabic-English, another advantage of NMT that is worth mentioning is its multilingual capabilities, making it much easier for translators to work across different languages, which is particularly beneficial for those working in multilingual projects or environments. Moreover, translators' roles can shift more towards post-editing and quality assurance where NMT systems can learn and adapt over time with human feedback, creating a positive feedback loop where human corrections and improvements enhance future translations.

Businesses and clients can benefit from the combined effort of NMT and human translators as it is more cost-effective. This combination enables the handling of large volumes of texts at a reduced cost while maintaining quality through human oversight. Previous research (Cui

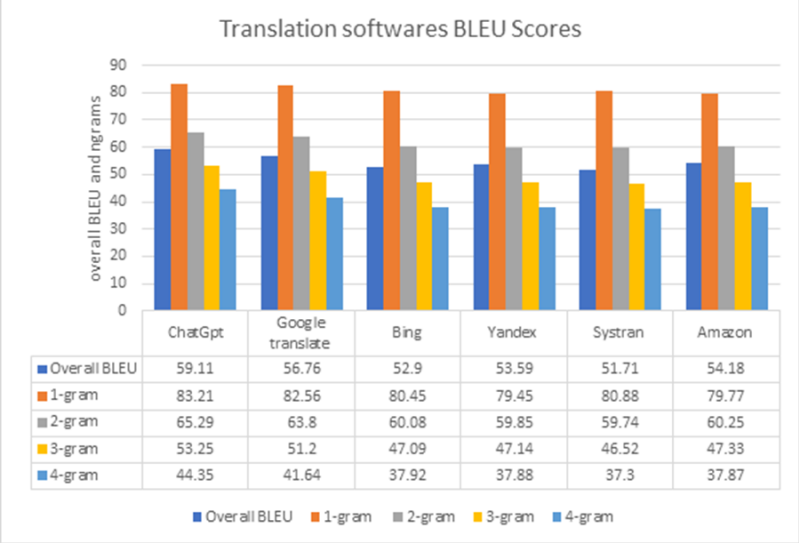
et. al. 2023; Wang & Daghigh, 2023) also suggests that NMT, when used in collaboration with human post-editing, can achieve quality levels like those of human-only translation efforts.

Based on this, the research findings indicate that human translators experience substantial benefits from employing NMT in their work. This aligns with the findings of Ameen and Ahmed (2023), whose study confirmed that adopting NMT in translation systems effectively simulates how the human brain operates in producing translations and learns from texts previously translated by human translators.

The findings revealed that ChatGPT-4, followed by Google Translate is the most reliable NMT system in translating general Arabic texts to English. This is measured by their performance in accuracy and fluency determined by the BLEU metric. The results also indicate different and multiple levels of effectiveness among and between all systems, as reflected in their BLEU scores. ChatGPT achieved the highest BLEU score which is (59.11), and this means that ChatGPT is considered the best system in terms of accuracy and fluency. Its advanced NMT algorithms effectively process complex sentences and idiomatic expressions, which results in translations that closely match the reference texts. Google Translate also performed very well in comparison to the other systems, with a BLEU score of 56.76, close to ChatGPT's score. Its strong contextual translation and its ability to generate natural-sounding translations contributed to its high score. Based on the n-grams, Google Translate shows effectiveness in capturing nuances and producing contextually appropriate translations. Similarly, Microsoft Bing's BLEU score of 52.9 means that it is an effective system in translation, though it is slightly less effective than ChatGPT-4 and Google Translate.

Amazon Translate's BLEU score of 54.18 indicated a very good performance, particularly in handling large volumes of text efficiently. However, its translations were sometimes less fluent than those of the top-performing systems, affecting its overall BLEU scores, especially in idiomatic and context-heavy content. Moreover, Yandex BLEU score is 53.59; a score that is lower than the previous systems. This means that Yandex translation includes frequent errors in grammar and in translating idiomatic expressions. While it performed quite well for simpler and easier sentences, it struggled with more complex linguistic structures, leading to lower overall effectiveness. Finally, Systran's BLEU score of 51.71 was the lowest among the evaluated systems. Although it handled technical translations, its translations lacked fluency and naturalness, resulting in the lowest BLEU score.

Figure 8 summarizes the BLEU scores for the six NMT systems, showing that ChatGPT has the highest overall BLEU score of 59.11, indicating it generally provides the most accurate translations. Google Translate follows with a score of 56.76, also demonstrating high translation quality. Amazon and Yandex have similar overall scores of 54.18 and 53.59, respectively. However, the results show that Bing and Systran have lower overall scores of 52.90 and 51.71, respectively.



	ChatGpt	Google translate	Bing	Yandex	Systran	Amazon
Overall BLEU	59.11	56.76	52.9	53.59	51.71	54.18
1-gram	83.21	82.56	80.45	79.45	80.88	79.77
2-gram	65.29	63.8	60.08	59.85	59.74	60.25
3-gram	53.25	51.2	47.09	47.14	46.52	47.33
4-gram	44.35	41.64	37.92	37.88	37.3	37.87

Figure 8. BLEU Scores for ChatGPT, Google MT, Bing MT, Yandex MT, Systran MT and Amazon

In terms of n-gram scores, the findings revealed that for 1-gram, which measures the accuracy of individual word translations, ChatGPT scores the highest at 83.21, followed closely by Google Translate at 82.56. Bing and Systran have scores above 80, while Amazon and Yandex are slightly below. For 2-gram, which measures the accuracy of pairs of consecutive words, ChatGPT again leads with a score of 65.29. Google Translate scores 63.80, followed by Amazon at 60.25, Bing at 60.08, Yandex at 59.85, and Systran at 59.74. For 3-gram, which measures the accuracy of triplets of consecutive words, ChatGPT has the highest score at 53.25, followed by Google Translate at 51.20, while other systems' scores are in the high 40s. Finally, for 4-gram, which measures the accuracy of quadruplets of consecutive words, ChatGPT scores the highest among all systems at 44.35, showing it maintains context well over longer sequences. Google Translate scores 41.64, with the rest scoring below 38.

There is a significant difference in the quality of translations among the tested systems. As for some other users who are in need for high quality translation of normal general text, ChatGPT-4 and Google Translate are highly recommended, and that is because of the advanced neural machine translation methods and technologies and the comprehensive languages models, which allow them to easily conduct and provide translations that are contextually correct.

These findings also match the results of the study of Khondaker et al. (2023), which shows that ChatGPT is a great accomplishment for direct texts, although it may require the intervention of human translators. In addition, the results of this study are also in line with those of Almahasees (2017), which determined that Google Translate surpasses Microsoft Bing when measured against approved human translations.

7. Conclusion

This study significantly contributes to the field of translation by investigating the capabilities of various neural machine translation (NMT) systems, focusing on the English-Arabic language pair, which presents unique linguistic and cultural challenges. However, limitations in this study include a restricted dataset of only 1,000 sentences, which may affect the generalizability of the findings. Additionally, the focus on English-Arabic translation may limit applicability to other language pairs, and reliance on a single evaluation tool, the BLEU metric, might not capture all aspects of translation quality, such as nativeness. Future research should consider a larger corpus across diverse text genres and incorporate human evaluations to enhance reliability (see e.g. Jarrar et al. 2024).

Moreover, the slight variations in scores among the NMT systems (ChatGPT: 59.11, Google Translate: 56.76, Amazon: 54.18, Yandex: 53.59) offer useful insights into their comparative performance. However, we must acknowledge that this conclusion is one of the limitations of this study as these differences may not be substantial enough to make definitive claims about overall translation quality. In light of these findings, it is recommended that users seeking high-quality translations utilize ChatGPT-4 and Google Translate due to their superior performance. Moreover, translation service providers should integrate NMT systems with human expertise, leveraging the strengths of both to enhance translation quality and efficiency. Future research could explore the performance of these systems in specialized fields and with less common language pairs to provide a more comprehensive understanding of their capabilities.

Acknowledgments

Not applicable.

Authors' contributions

Rand Habib was involved in the conception and design of the study, as well as the analysis and interpretation of the data. Additionally, she contributed substantially to the drafting of the paper and revising it critically for intellectual content. Prof. Linda Alkhawaja contributed to the conception and design of the study, along with the analysis and interpretation of the data. Furthermore, she played a significant role in drafting the paper and revising it critically for intellectual content. Dr. Ogareet Khoury and Dr. Sa'ida Al-Sayyed participated in the analysis and interpretation of the data, providing valuable insights. They also contributed to the drafting of the paper and revising it critically for intellectual content. All authors have given final approval for the version to be published and agree to be accountable for all aspects of the work.

Funding

Not applicable.

Competing interests

Not applicable.

Informed consent

Not applicable.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Abanomey, A. A., Almossa, S. Y. (2023). Translation quality assessment practices of faculty members of colleges of languages and translation in Arab countries: an exploratory study. *Humanit Soc Sci Commun, 10*, 835. https://doi.org/10.1057/s41599-023-02352-z

Abdelaal, N. M., & Alazzawie, A. (2020). Machine translation: the case of Arabic-English translation of news texts. *Theory and Practice in Language Studies, 10*(4), 408. https://doi.org/10.17507/tpls.1004.09

Abu Manie, H. I., Al-Sayyed, S. W., Alkhawaja, L. A., & Rababa'h, B. B. (2025). Congratulation strategies of Crown Prince Hussein's wedding: A socio-pragmatic study of Facebook comments. *Open Cultural Studies*, *9*(1), Article 20250051. https://doi.org/10.1515/culture-2025-0051

Adawiyah, A. R., Baharuddin, B., Wardana, L. A., & Farmasari, S. (2023). Comparing post-editing translations by Google NMT and Yandex NMT. *Teknosastik (Bandar Lampung)*, *21*(1), 23. https://doi.org/10.33365/ts.v21i1.2339

Ahmed, A., & Lenchuk, I. (2024). The interaction between morphosyntactic features and the performance of machine translation tools: the case of Google Translate, Systran, and Microsoft Bing in English-Arabic translation. *Theory and Practice in Language Studies*, *14*(2), 614-625. https://doi.org/10.17507/tpls.1402.35

AlIssa, A. (2024). The Effectiveness of Employing Jordanian School Teachers the Artificial Intelligence Applications in Blended Learning. *Al-Balqa Journal for Research and Studies*, *27*(4), 39-55. https://doi.org/10.35875/827jzc89

Alkatheery, E. R. (2023). Google Translate Errors in Legal Texts: Machine Translation Quality Assessment. *Arab World English Journal for Translation & Literary Studies*, *7*(1), 208-219. https://doi.org/10.24093/awejtls/vol7no1.16

Alkhawaja, L. (2023). Artificial intelligence in education: Harnessing its power as a valuable tool, not an adversary. *International Journal of Computer-Assisted Language Learning and Teaching*, *13*(1), Article e329607. https://doi.org/10.4018/IJCALLT.329607

Alkhawaja, L. (2024). Unveiling the new frontier: ChatGPT-3 powered translation for Arabic-English language pairs. *Theory and Practice in Language Studies*, *14*(2), 347-357. https://doi.org/10.17507/tpls.1402.05

Almahasees, Z. (2017). Assessing the Translation of Google and Microsoft Bing in Translating Political Texts from Arabic into English. *International Journal of Languages, Literature and Linguistics (Print)*, *3*(1), 1-4. https://doi.org/10.18178/ijlll.2017.3.1.100

Almahasees, Z. (2021). *Analysing English-Arabic Machine Translation: Google Translate, Microsoft Translator and Sakhr.* Routledge. https://doi.org/10.4324/9781003191018

Al-Sabbagh, R. (2023). The Negative Transfer Effect on the Neural Machine Translation of Egyptian Arabic Adjuncts into English: The Case of Google Translate. *International Journal of Arabic-English Studies.* https://doi.org/10.33806/ijaes.v24i1.560

Ameen, H. H. M., & Ahmed, H. A. (2023). Assessing the Quality of Machine Translation from Kurmanji Kurdish into English. *Academic Journal of Nawroz University*, *12*(3), 503-517. https://doi.org/10.25007/ajnu.v12n3a1690

Ashraf, R. (2023). Demystifying the BLEU Metric: A Comprehensive Guide to Machine Translation Evaluation | TraceLoop blog. Retrieved from https://www.traceloop.com/blog/demystifying-the-bleu-metric

Beseiso, M., Tripathi, S., Al-Shboul, B., & Aljadid, R. (2022). Semantics-based English-Arabic machine translation evaluation. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(1), 189-197. https://doi.org/10.11591/ijeecs.v27.i1.pp189-197

Callison-Burch, C. (2009). *Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk.* ResearchGate. https://doi.org/10.3115/1699510.1699548

Chen, J. (2022). Analysis of intelligent translation systems and evaluation systems for Business English. *Journal of Mathematics* (Hindawi. Print), 2022, 1-7. https://doi.org/10.1155/2022/5952987

Cui, Y., Liu, X., & Cheng, Y. (2023). A Comparative Study on the Effort of Human Translation and Post-Editing in Relation to Text Types: An Eye-Tracking and Key-Logging Experiment. *Sage Open*, *13*(1). https://doi.org/10.1177/21582440231155849

Dalibor, D. (2024). *Translation Software: What it is, and how to choose the best one | phrase*. Phrase. Retrieved from https://phrase.com/blog/posts/translation-software/

Datta, G., Joshi, N., & Gupta, K. (2022). Analysis of Automatic Evaluation Metric on Low-Resourced Language: BERTScore vs BLEU Score. In *Lecture notes in computer science* (pp. 155–162). https://doi.org/10.1007/978-3-031-20980-2_14

De Oliveira, R. G., & Anastasiou, D. (2011). Comparison of SYSTRAN and Google Translate for English→Portuguese. *Revista Tradumàtica*, *9*, 118-136. https://doi.org/10.5565/rev/tradumatica.14

Diab, N. (2021). Out of the BLEU: An Error Analysis of Statistical and Neural Machine Translation of WikiHow Articles from English into Arabic. *CDELT Occasional Papers: In the Development of English Language Education (Print)*, *75*(1), 181-211.

https://doi.org/10.21608/opde.2021.208437

Dorr, B., Snover, M., & Madnani. (2011). Chapter 5: Machine Translation Evaluation. In Dorr, B., Olive, J., McCary, J., &Christianson, C. (2011). Machine translation evaluation and optimization. In Handbook of natural language processing and machine translation (pp. 745-843). *Springer*, New York, NY. https://doi.org/10.1007/978-1-4419-7713-7_5

Fernandes, P., Farinhas, A., Rei, R., De Souza, J. G. C., Ogayo, P., Neubig, G., & Martins, A. F. T. (2022). *Quality-Aware decoding for neural machine translation. arXiv* (Cornell University). https://doi.org/10.18653/v1/2022.naacl-main.100

Gala, D., & Makaryus, A. N. (2023). The Utility of Language Models in Cardiology: A narrative review of the benefits and concerns of CHATGPT-4. *International Journal of Environmental Research and Public Health (Online)*, *20*(15), 6438. https://doi.org/10.3390/ijerph20156438

Gaurav, M., Saikumar, G., Srivastava, A., Natarajan, P., Ananthakrishnan, S., & Matsoukas, S. (2013). Leveraging Arabic-English Bilingual Corpora with Crowd Sourcing-Based Annotation for Arabic-Hebrew SMT. *In Lecture Notes in Computer Science* (pp. 297-310). https://doi.org/10.1007/978-3-642-37256-8_25

Ghosh, S., & Caliskan, A. (2023). ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *ACM Digital Library*. https://doi.org/10.1145/3600211.3604672

Ismailia, T. (2023). Analysis of Machine Translation Performance on Translating Informative Text from English into Indonesian. *Ebony*, *3*(2), 129-138. https://doi.org/10.37304/ebony.v3i2.9809

Jabak, O. (2019). Assessment of Arabic-English translation produced by Google Translate. *International Journal of Linguistics, Literature & Translation*, *2*(4), 10.

Jarrar, Y. A., Al-Sayyed, S. I. W., Rababa'h, B. B., & Alkhawaja, L. A. (2024). A corpus-based analysis of four near-synonymous English verbs. *Forum for Linguistic Studies*, *6*(6), 678-701. https://doi.org/10.30564/fls.v6i6.7378

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). *Is ChatGPT a good translator? Yes, with GPT-4 as the engine.* arXiv (Cornell University). https://doi.org/10.48550/arxiv.2301.08745

Kadaoui, K., Magdy, S. M., Waheed, A., Khondaker, M. T. I., El-Shangiti, A. O., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). TARJAMAT: Evaluation of BARD and ChatGPT on machine translation of ten Arabic varieties. arXiv (Cornell University). https://doi.org/10.18653/v1/2023.arabicnlp-1.6

Kalinina, M. V., & Kalinina, M. (2023). Conflictogenicity Of Digital Communication Forms in The Aspect Of Linguistic Security (on the comments on blogs of the Yandex-Zen platform). *Vestnik Baltijskogo Federal'nogo Universiteta Im. I. Kanta*, *2*, 17-27. https://doi.org/10.5922/pikbfu-2023-2-2

Keary, T. (2024). Amazon Translate. *Techopedia*. Retrieved from https://www.techopedia.com/definition/amazon-translate

Kenny, D. (2022). *Human and machine translation.* In Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence, Ed. Dorothy Kenny. Language Science Press: Germany.

Khondaker, M. T. I., Waheed, A., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). *GPTARAEVaL: A Comprehensive Evaluation of ChatGPT on Arabic NLP.* arXiv (Cornell University). https://doi.org/10.18653/v1/2023.emnlp-main.16

Klímová, B., Pikhart, M., Benites, A. D., Lehr, C., & Sanchez-Stockhammer, C. (2022). Neural machine translation in foreign language teaching and learning: a systematic review. *Education and Information Technologies*, *28*(1), 663-682. https://doi.org/10.1007/s10639-022-11194-2

Lengkong, O., Mandias, G. F., & Tombeng, M. T. (2022). The implementation of Yandex Engine on Live Translator application for Bahasa and English using Block Programming MIT App Inventor Mobile based. *Cogito Smart Journal* (Online), *8*(1), 92-101. https://doi.org/10.31154/cogito.v8i1.388.92-101

Marie, B., Fujita, A., & Rubino, R. (2021). *Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 7297-7306). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.566

Mart ń, B. S. (2017). *Translation Quality assessment of Google Translate and Microsoft Bing Translator.* Retrieved from https://uvadoc.uva.es/handle/10324/22596

Maučec, M., & Donaj, G. (2019). *Machine Translation and the Evaluation of Its Quality*. In Natural Language Processing - New Approaches and Recent Applications.

Moneus, A. M. A., & Sahari, Y. (2024). Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon* (Londen), e28106. https://doi.org/10.1016/j.heliyon.2024.e28106

Musaad, M., & Towity, A. (2023). Translation Evaluation of Three Machine Translation Systems, with Special References to Idiomatic Expressions. *Mağallaẗ Al-ʿulūm Al-tarbawiyyaẗ Wa-al-dirāsāt Al-insāniyyaẗ Silsilaẗ Al-ādāb Wa-al-ʿulūm Al-tarbawiyyaẗ*

Wa-al-insāniyyaẗ Wa-al-taṭbīqiyyaẗ, 29, 678-708. https://doi.org/10.55074/hesj.vi29.700

Ni'mah, I., Fang, M., Menkovski, V., & Pechenizkiy, M. (2023). *NLG Evaluation Metrics Beyond Correlation Analysis: An Empirical Metric Preference Checklist.* NLG Evaluation Metrics Beyond Correlation Analysis: An Empirical Metric Preference Checklist. https://doi.org/10.18653/v1/2023.acl-long.69

Peng, K., Liu, D., Zhong, Q., Shen, L., Liu, X., Zhang, M., … Tao, D. (2023). Towards making the most of ChatGPT for machine translation. *Social Science Research Network.* https://doi.org/10.2139/ssrn.4390455

Popel, M., Tomkova, J., Tomek, Ł., Kaiser, J., Uszkoreit, O., Bojar, et al. (2020). Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications, 11*(1). https://doi.org/10.1038/s41467-020-18073-9

Putri, R. S. Y., Sofyan, R., & Nasution, E. H. (2022). Translation quality assessment on Medan City Tourism official web pages. *LingPoet, 3*(1), 32-39. https://doi.org/10.32734/lingpoet.v3i1.6500

Rescigno, A. A., Monti, J., Way, A., & Vanmassenhove, E. (2020). *A case study of natural gender phenomena in translation. A comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish.* In Accademia University Press eBooks (pp. 359-364). https://doi.org/10.4000/books.aaccademia.8844

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation, 56*(2), 593-619. https://doi.org/10.1007/s10579-021-09537-5

Rossi, C., & Carré, A. (2022). How to choose a suitable neural machine translation solution. In D. Kenny (Ed.), *Machine translation for everyone: empowering users in the age of artificial intelligence.* Language Science Press: Germany.

Sallam, M., & Mousa, D. (2024). Evaluating ChatGPT performance in Arabic dialects: A comparative study showing defects in responding to Jordanian and Tunisian general health prompts. *Mesopotamian Journal of Artificial Intelligence in Healthcare*, *2024*, 1-7. https://doi.org/10.58496/mjaih/2024/001

Sanz-Valdivieso, L., & Arroyo, B. L. (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

Shalabi, A., & Abu Amrieh, Y. (2025). Rawi Hage's Cockroach and Laila Lalami's The Other Americans: Images of twenty-first century Occident in Arab eyes. *Textual Practice*, *39*(2), 261-283. https://doi.org/10.1080/0950236X.2023.2288115

Shalabi, A., & Amrieh, Y. A. (2024). Occidentalism revisited: Insights from contemporary Anglophone Arab diasporic literature. *Journal of Intercultural Communication*, *24*(4), 134-145. https://doi.org/10.36923/jicc.v24i4.977

Shukla, A., Bansal, C., Badhe, S., Ranjan, M., & Chandra, R. (2023). *An evaluation of Google Translate for Sanskrit to English translation via sentiment and semantic analysis. arXiv* (Cornell University). https://doi.org/10.1016/j.nlp.2023.100025

Sismat, M. a. H. (2022). Analyzing patterns of errors in neural and statistical machine translation of Arabic and English. *JALL | Journal of Arabic Linguistics and Literature*, *2*(2), 126-142. https://doi.org/10.59202/jall.v2i2.347

Tatoeba. (2024). *Tatoeba is a collection of sentences and translations.* Retrieved from https://tatoeba.org/en/

Teebagy, S., Colwell, L., Wood, E. R., Yaghy, A., & Faustina, M. (2023). Improved Performance of ChatGPT-4 on the OKAP Exam: A Comparative Study with ChatGPT-3.5. *medRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.04.03.23287957

Vanjani, M., & Aiken, M. (2020). A Comparison of Free Online Machine Language Translators. *Journal of Management Science and Business Intelligence, 5*(1), 26-31. https://doi.org/10.5281/zenodo.3961085

Wang, J., & Gu, Q. (2024). Quantitative Assessment of Translation Quality in Education, Certification, and Industry: An Overview. *Open Journal of Modern Linguistics*, *14*, 209-223. https://doi.org/10.4236/ojml.2024.142012

Wang, Y., & Daghigh, A. J. (2023). Effect of text type on translation effort in human translation and neural machine translation post-editing processes: evidence from eye-tracking and keyboard-logging. *Perspectives*, 1-16. https://doi.org/10.1080/0907676X.2023.2219850

Warner, A. 2022. *Humans still beat machines when it comes to literary translation.* Retrieved from https://multilingual.com/literary-machine-translation/

Xiao, Y., Wu, L., Guo, J., Li, J., Zhang, M., Qin, T., & Liu, T. (2023). A Survey on Non-Autoregressive Generation for Neural Machine Translation and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. https://doi.org/10.1109/tpami.2023.3277122

Yaseen, M., Al-Sayyed, S., & Ibnian, S. (2024). Exploring Vocabulary Development and Student Preferences: A Comparative Study of Digital and Print Extensive Reading in an EFL Context. *Al-Balqa Journal for Research and Studies, 27*(4), 1-18. https://doi.org/10.35875/r2p7np21

Yin, Y., Li, Y., Meng, F., Zhou, J., & Zhang, Y. (2023). *Categorizing semantic representations for neural machine translation.* arXiv (Cornell University). https://doi.org/10.48550/arxiv.2210.06709

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic Machine Translation: A Survey with Challenges and Future Directions. *IEEE*. https://doi.org/10.1109/ACCESS.2021.3132488

Zhivotova, A. A., Berdonosov, V. D., & Redkolis, E. V. (2020). Improving the quality of scientific articles machine translation while writing original text. *International Multi-Conference on Industrial Engineering and Modern Technologies* (FarEastCon). https://doi.org/10.1109/fareastcon50210.2020.9271442

Zughoul, M. R., & Abu-Alshaar, A. M. (2005). English/Arabic/English Machine Translation: A Historical perspective. *Meta*, *50*(3), 1022-1041. https://doi.org/10.7202/011612ar