



A Small-Scale Comparative Evaluation of ChatGPT-3.5 and Google Translate for English-Urdu and Urdu-English Translation

Aatka Faryal Riaz^{1*}, Rafay Azmat², and Rubeka Sehar³

¹Department of Computer Science, International University of Applied Science,
Bad Honnef, Germany

²Department of Electronics Engineering, Mehran University of Engineering & Technology,
Jamshoro, Pakistan

³ICT Department, University of Sindh, Jamshoro, Pakistan

Abstract: Today, technologies like ChatGPT and Google Translate are utilized for translating text and enabling smooth exchange of information. Urdu is the national language of Pakistan, while English serves as the official language and the primary medium of communication in schools and workplaces. Since people commonly speak Urdu, they often use machine translators for English to Urdu and Urdu to English translation. These machine translators can produce translations with a diverse vocabulary but sometimes their translations are not contextually correct or fully reliable. Therefore, it is essential to evaluate the translation outputs of the machine translators. The prime objective of this paper is to compare the performance of machine translation systems at the paragraph level, enabling users to identify the most effective one. In this research, Google Translate and ChatGPT-3.5 were used as translation tools to translate text between English and Urdu in both directions. The texts were selected from different categories, such as historical texts, poetry, literary works, novels, and legal documents. The machine translation systems translated English text into Urdu, and the generated outputs were compared and assessed by authors. Similarly, Urdu text was translated into English, and the resulting translations were compared and assessed by authors. We assessed and compared the translated outputs based on phrasing style, linguistic accuracy, and contextual meaning. To ensure a reliable comparison, we calculated the BiLingual Evaluation Understudy (BLEU) and Metric for Evaluation of Translation with Explicit ORdering (METEOR) scores of the translations generated by these translation tools. Based on our analysis and BLEU and METEOR scores, we concluded that the performance of ChatGPT-3.5 is better than Google Translate.

Keywords: Machine Translators, Chat GPT, Google Translate, English-Urdu Language, Urdu-English Language, Evaluation Metrics, BLEU, METEOR.

1. INTRODUCTION

Pakistan is a multilingual country where numerous languages are spoken as native tongues. Urdu is the national language of Pakistan, and the official language of Pakistan is English. According to Ethnologue [1], Urdu ranks as the eleventh most extensively spoken language globally, with a total of 246 million speakers including those who use it as a secondary language and approximately 1.5 billion people around the world speak English. According to international center for language studies [2], even

though English is the most widely spoken language globally, only 25% are native speakers, while the rest have acquired it as a second language. Almost all online media is in English, with over half of all websites (52.1%) using it for their content.

As English is the official language, and school textbooks are written in English. However, since students primarily speak Urdu, they often rely on machine translators to write essays or assignments in English language. They also frequently use these tools to comprehend English-language content.

To gain a deeper understanding, they translate entire paragraphs rather than isolated sentences. This research evaluates the accuracy of machine translators at the paragraph level, as opposed to single-sentence translations, to better reflect the above requirement. Moreover, nowadays people from Pakistan are relocating and migrating to other countries for their jobs and business. Students are travelling to other countries for their studies. They are required to understand and communicate in different languages, so they have to use translators. Owing to the need for instant translations, they use machine translators for effective exchange of information. The translation generated by these machine translators is of high quality and employs diverse synonyms; however, it is not always correct and dependable. There is a need to assess the accuracy of their generated translations. Once users are aware of the accuracy levels, it will be easy for them to choose the most reliable translator.

This research aims to conduct a comparative evaluation of the translation outputs produced by ChatGPT-3.5 and Google Translate to help people to select the best one. The ChatGPT-3.5 model and Google Translate were used in this research. We opted for ChatGPT due to the significant attention it has recently received, while Google Translate was chosen because of its widespread popularity. To evaluate we gave a set of English paragraphs as input to these translators and compare the quality of their outputs (i.e., Urdu translations) using evaluation metric BiLingual Evaluation Understudy (BLEU) and Metric for Evaluation of Translation with Explicit Ordering (METEOR). Similarly, we checked their performance for Urdu to English translation. In literature, most of the research to evaluate the performance of machine translators has focused on using sentences rather than paragraphs. For example, Malik and Habib [3] used seven different sentences, not paragraphs, to assess machine translators in their study. Similarly, Ghassemiazghandi [4] evaluated ChatGPT's performance for Persian-to-English translation using sentence-level assessment. In this study, we mainly focused on using paragraphs instead of individual sentences. We conducted a comparative evaluation of translations produced by ChatGPT-3.5 and Google Translate with respect to contextual relevance, phrasing style, and word choice using expert human translations as a benchmark. To validate our analysis, we also calculated the BLEU

and METEOR scores of each generated translation to determine which system performed better. To the best of our knowledge, this work is unique in evaluating and comparing the translation quality of paragraphs rather than individual sentences for English to Urdu and Urdu to English translations.

2. MATERIALS AND METHODS

We used ChatGPT and Google Translate in this research. These state-of-the-art technologies are described in this section.

2.1. ChatGPT

According to Biswas [5], due to the fast progress in the field of Artificial Intelligence (AI) and Natural Language Processing (NLP), highly sophisticated language models have been developed, such as ChatGPT, which is a large language model developed by OpenAI. This model employs deep learning techniques to interpret and generate text that is very similar to the human communication. ChatGPT is built using transformer architecture, which enables it to produce responses in a conversational manner. As it can produce text in a human style, it holds significant potential to impact areas such as healthcare, research, and language translation, as explained by Biswas [6].

2.1.1. A transformer-based architecture

ChatGPT (Generative Pretrained Transformer) was created to overcome the constraints of previous language processing techniques, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), as discussed by Ray [7]. ChatGPT utilizes transformer models, which allow it to recognize patterns and grammatical structures of languages. The transformer is a neural network framework that employs a self-attention mechanism. This architecture has two principal advantages. First, it is highly effective at applying context in language processing tasks. For example, when a sentence is processed, the meaning of a word or phrase may vary considerably depending on its context, which can be difficult to capture by using only conventional grammatical rules because it also relies on the functional relationships between the words. The self-attention mechanism in the transformer, however, enables the model to incorporate contextual information in a data-

driven manner. Second, the transformers permit substantial parallelization and can attain cutting-edge translation accuracy after just 3.5 days of training on eight GPUs, as investigated by Vaswani *et al.* [8].

2.1.2. How ChatGPT is trained

ChatGPT is trained on massive volumes of text data, including web content, articles, and books, enabling it to generate grammatically correct and contextually appropriate output. A brief overview of ChatGPT's training process is presented here. According to Memon [9], the model's training occurs in the following three stages.

1. Generative Pre-Training
2. Supervised Fine-Tuning (SFT)
3. Reinforcement Learning through Human Feedback (RLHF)

Stage 1: Generative Pre-Training

At the first stage, the model is trained on a vast collection of textual data gathered from across the internet, including content from books, magazines, articles, and other sources. The dataset covers a wide range of genres and subjects, enabling the model to produce text in diverse linguistic forms and contextual backgrounds. Stage 1 equips the model with general language capabilities such as summarizing text, completing sentences, translating content, and analyzing sentiment, but it is not specialized in doing a particular task. For instance, the model performs well in summarizing lengthy articles into concise versions. However, at this stage, it cannot engage in topic-specific conversations with humans, even though it excels in general language processing.

Stage 2: Supervised Fine-Tuning

Supervised Fine-Tuning (SFT) represents the 2nd stage of ChatGPT's training. In this phase, the model is trained to perform specific tasks, such as engaging in conversational dialogue. Until this point, ChatGPT's knowledge is only limited to the data on which it has been trained. You can think of it as having a small snapshot of the world in its "brain." So, if you ask it something outside of what it has learned, it may give a wrong or any unpredictable answer.

Stage 3: Reinforcement Learning through Human Feedback (RLHF)

This is the third stage of training. In this stage, the model undergoes additional training to respond to queries it was not explicitly taught. This stage uses reinforcement learning, where the model is guided by a reward system and receives feedback from humans to improve its responses. During the third stage of training, a human agent was also involved to guide the process of reinforcement learning. This is why it is referred to as Reinforcement Learning with Human Feedback (RLHF).

2.1.3. How ChatGPT performs language translation

ChatGPT is not intended solely for translation, but its ability to follow complicated instructions makes it a valuable tool for translation, as noted by Siu [10]. ChatGPT can facilitate seamless interaction between people communicating in different languages. For example, if a user's primary language is Urdu, he can enter text in Urdu, and ChatGPT will generate an immediate translation in the target language. Below is a concise description of ChatGPT's translation process.

As described earlier, the training of ChatGPT consists of three stages. The pre-training and fine-tuning phases equip the model with language translation abilities. In the pre-training phase, the model learns from a large collection of multilingual data, which helps it to understand general linguistic patterns. Then, in the fine-tuning stage, ChatGPT is trained using aligned sentence pairs in different languages, such as English and Urdu, to perform specific translation tasks. The model then learns to generate translations by predicting the most appropriate output sentence based on the input. While its translations are generally accurate, it occasionally struggles with complex texts, such as literary content.

2.2. Google Translate

Google Translate is a machine translation service which is developed by Google to provide translations between different languages. It enables direct translation of text from one language to another, as discussed by Boitet *et al.* [11]. As of 2022, Google Translate provided a free automated

translation service for 133 languages [12], including Urdu. The translation service uses Neural Machine Translation (NMT) framework. Furthermore, it is widely used, with over 500 million daily users and approximately 100 billion words translated each day [13]. The widespread use of Google Translate has led to a decline in the use of traditional digital dictionaries, the researchers who study English as a Foreign Language (EFL) course at a Chilean high school suggested to employ Google Translate to assist students in enhancing their syntax and grammar, as discussed by Chang *et al.* [14]. Similarly, Tsai's study [15] also emphasized the effectiveness of Google Translate in reducing typing errors and producing longer texts.

2.2.1. Google translate architecture

As explained by King [16], Google Translate's translation system processes information in a manner similar to the human brain: it ingests data, encodes it, situates it within a context, and then re-encodes it to produce an output. The Google's Neural Machine Translation (GNMT) system architecture includes an encoder, a decoder, and an attention mechanism. In the encoder, the bottom layer operates bidirectionally, with the green nodes processing words from right to left and the pink nodes processing words from left to right. To accelerate training, the model is distributed across multiple GPUs. The system has eight Long Short-Term Memory (LSTM) layers each in the encoder and decoder, as illustrated in Figure 1 by Wu *et al.* [17].

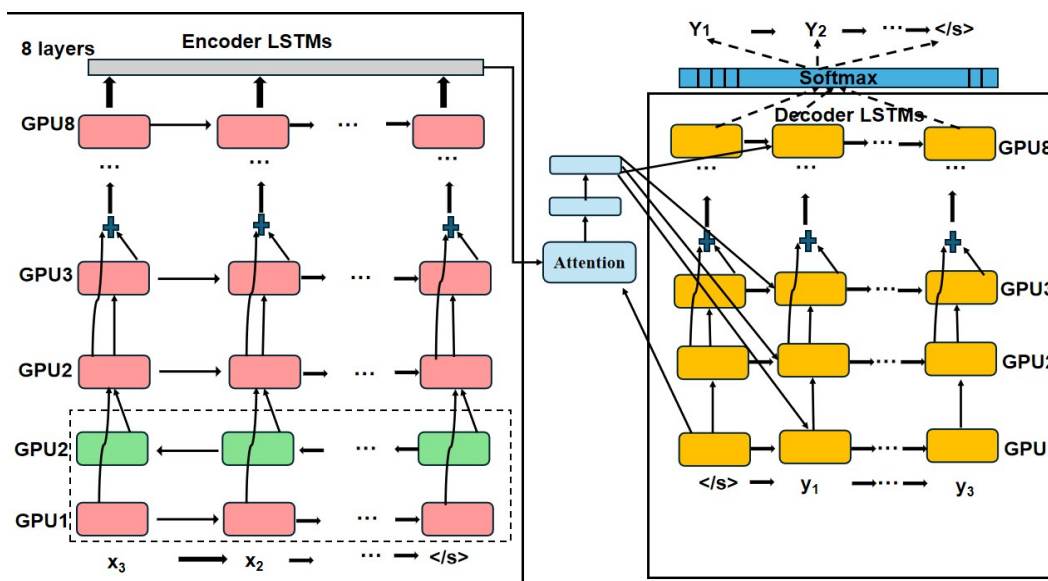


Fig. 1. Model architecture of Google Translate.

2.2.2. Working principle of Google Neural Machine Translation

The Google's Neural Machine Translation (GNMT) system uses its encoder, decoder, and attention module to perform language translation. For example, translating a sentence from English to Urdu involves the following steps, as illustrated in Figure 2.

1. The English sentence is fed into the encoder, where it is converted into a fixed-length vector that can only be understood by a computer.
2. Then the attention mechanism then determines which English words should be assigned greater weight to produce the corresponding Urdu words.
3. Finally, the decoder generates the Urdu translation one word at a time, by focusing its attention on the words determined by the attention mechanism.

2.3. Research Methodology

For the present study, we selected 14 different paragraphs and 2 idiomatic sentences to assess and compare the accuracy of translation of ChatGPT and Google Translate. In order to generalize the results, we selected paragraphs from different genres, for instance, historical texts, poetry, literary and religious writings, and legal texts. Each paragraph consists of approximately 3 to 4 sentences. We took texts from history, poetry, novels, and literature because English in these genres is often complex.

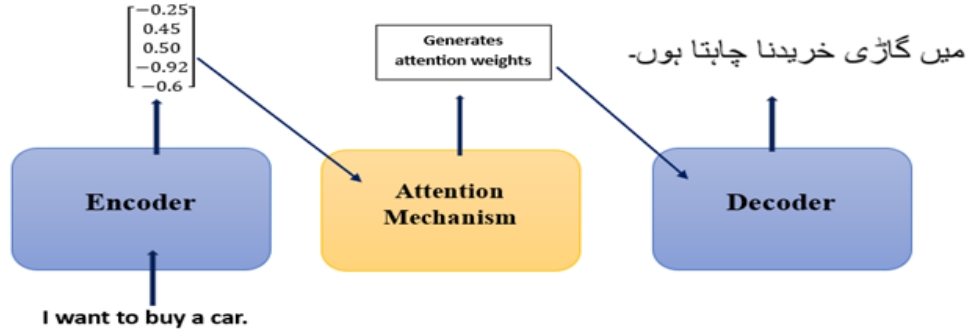


Fig. 2. The language translation process of the Google Translate.

In contrast, we selected texts from articles, textbooks, and newspapers because they usually contain simpler language. All texts have been collected from reliable sources such as textbooks (e.g., Punjab Textbook Board), published novels, and newspaper articles. Although these machine translators can translate English language into several other languages, but we mainly focused on English and Urdu languages because of the large number of English language users in Pakistan who rely on these tools for various purposes.

The methodology used in this research is illustrated in Figure 3. English text was input into ChatGPT-3.5 and Google Translate to obtain Urdu translations. The BLEU and METEOR scores were calculated for each translation. The authors then assessed the accuracy of generated translations with respect to context, selection of words, and phrasing style, in comparison with

professional human translations and compared the BLEU and METEOR scores to determine which system performed better. The same approach was applied to assess the accuracy of Urdu-to-English translations. We translated each paragraph multiple times with both ChatGPT-3.5 and Google Translate and used the more accurate translation for analysis.

2.4. Dataset

The dataset consists of 14 different paragraphs and 2 idiomatic sentences. These paragraphs were selected from various genres, such as history, poetry, etc. We also included human translations for each paragraph to use as reference texts for evaluating the translation quality of machine translation outputs and for calculating BLEU and METEOR scores. These human translations were carefully prepared by professional translators who are proficient in the target language.

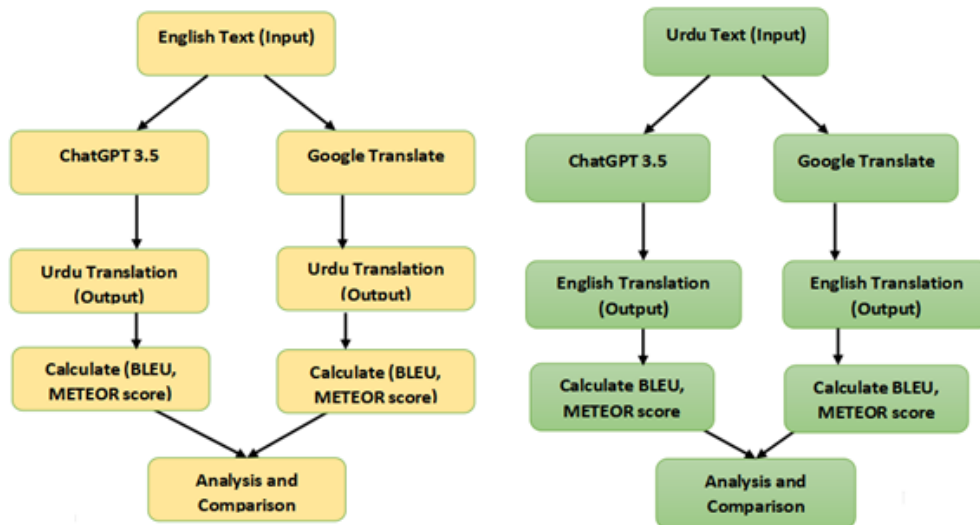


Fig. 3. Layout of the research methodology.

2.5. Performance Metric

As we are assessing the translations of paragraphs which consist of several lines. In addition to human observation, we incorporated an automatic evaluation method to enhance the credibility and reliability of machine translation assessment. Specifically, we used BLEU and METEOR as an evaluation metric to compare the accuracy of machine translators for English to Urdu translation and vice versa. According to the rule, the machine translator that achieved higher METEOR and BLEU score will be considered more accurate. This is the essence of the evaluation of machine translators.

2.5.1. BLEU

BLEU stands for BiLingual Evaluation Understudy. The BLEU is an algorithm to assess the quality of translated text that is translated from one natural language to another. According to Papineni *et al.* [18], the core idea behind BLEU is the quality of the translation increases when it looks more similar to the professional human translation. The output value of BLEU always lies between 0 and 1. This value indicates the degree of similarity between the reference translation (i.e., human translation) and candidate sentences (i.e., machine translator output). The machine translation that exhibits a higher resemblance to the reference translation got value closer to 1. A higher BLEU score shows better quality of machine translation system as described by Ayesha *et al.* [19].

2.5.1.1. N-grams in bleu score:

To calculate the BLEU score we used NLTK (Natural Language Toolkit), a popular open-source library for Natural Language Processing (NLP) in Python, which allows to set the value of n-grams in the calculation of the BLEU score. Here 'n' is the set of consecutive words in a sentence. We can have n-grams as unigram (1-gram), bigram (2-gram), and trigram (3-gram) depending on the length of sentence. For example, for a sentence “**The water is colorless**”, we can have n-grams as:

- **1-gram (unigram):** The unigram approach focuses only on individual word matches during comparison between the machine translation and human translation, e.g., “The”, “water”, “is”, “colorless”.

- **2-gram (bigram):** The bigram approach considers matches of two consecutive words during comparison between the machine translation and human translation, for example, “The water”, “water is”, “is colorless”.
- **3-gram (trigram):** The trigram approach considers matches of three consecutive words during comparison between the machine translation and human translation, for example, “The water is”, “water is colorless”.
- **4-gram:** The 4-gram approach considers matches of four consecutive words, for example, “The water is colorless”.

2.5.1.2. How bleu score is calculated:

BLEU uses a modified form of precision to measure how closely the predicted text matches the reference text. The BLEU precision score can be determined using the following formula,

$$P = m_{max} / W_t$$

Where P is precision value of BLEU, m_{max} is the number of words found similar in reference sentence and predicted text and W_t is total number of words in predicted sentence.

To explain how BLEU score is calculated assume two sentences. A predicted text generated by a machine translator and a reference text which is a human translated text.

- **Reference Text:** Our guest arrived late because it was raining
- **Predicted Text:** Our guest arrived late because of the rain

The unigram calculation focuses on individual word match. In the following example, out of 8 words in the predicted text 5 words are correctly matched with the reference text. So, the total similar words (m_{max}) between predicted sentence and reference sentences are 5, and the number of words in predicted sentence (W_t) is 8. Thus, using the BLUE formula gives us the precision value of 0.625 for this predicted sentence.

Example:

Reference Text: Our guest arrived late because it was raining.



Predicted Text: Our guest arrived late because of the rain.

Here in our code, we used a unigram and bigram approach with smoothing function to get a more reliable BLEU score when the predicted text of a machine translator is short or has missing n-grams.

2.5.2. Meteor

METEOR stands for Metric for Evaluation of Translation with Explicit Ordering, is a metric used to measure the quality of translation generated by machine translators. The METEOR was developed at the Carnegie Mellon University, Pittsburgh, USA as described by Banerjee and Lavie [20]. Unlike BLEU, which primarily focuses on Precision, the METEOR is the more balanced approach as it considers both PRECISION AND RECALL. Recall evaluates how effectively the candidate translation includes all the words or information from the reference translation. This comprehensive approach enables METEOR to offer a more well-rounded assessment of translation quality.

One of the key strengths of the METEOR score is its ability to look beyond exact word matches. It includes stemming, which simplifies words to their base forms (e.g., “clapping” becomes “clap”), and synonym recognition (e.g., “quick” and “fast”), which treats similar words as valid matches. This helps METEOR better handle differences in language, such as varying word forms or different ways of saying the same thing. Assume the following reference translation **Our guest arrived late because it was raining** and candidate translation **Our guest came late because of the rain**.

- **Exact Match:** “Our”, “guest”, “late”, “because”
- **Stemming:** “raining” and “rain”
- **Synonym:** “arrived” and “came”

The METEOR aligns these matches and then calculates the score that reflects the quality of translation. We calculated the METEOR score by using Natural Language Toolkit (NLTK), a popular open-source library for Natural Language Processing (NLP) in Python.

3. RESULTS AND DISCUSSION

As we discussed, we used different types of paragraphs in this research to assess the translation quality of machine translation systems. In this section, we included one example of Urdu to English translation and one example of English to Urdu translation in tabular form (other examples are given in supplementary file, Tables S1-S14). Each table presents the selected text in its original form, the translation generated by ChatGPT, the translation produced by Google Translate, BLEU

and METEOR scores of both machines, and the translation by a professional human. In this section we presented our analysis and comparison about the accuracy of translations generated by ChatGPT-3.5 and Google Translate. We also validate our analysis with the help of the BLEU and METEOR score of both translators.

The text presented in Table 1 is an example of Urdu to English translation taken from the Pakistan studies book of Punjab textbook board [21]. According to our analysis the translation generated by ChatGPT is correct, however, the translation of the last sentence produced by Google Translate “Humans are obligated to enter Islam and perform the duties that Allah has imposed on them” is incorrect. This translation shows that humans are forced to enter Islam and perform religious duties, whereas the correct meaning should be humans are bound to pay duties of Allah after they have entered Islam. This intended meaning is also reflected in the human translation provided by professional translator. The Google Translate’s version appears to force humans to enter Islam which is not true. Moreover, if we compare the BLEU and METEOR scores of ChatGPT and Google Translate presented in Table 1, we noticed that ChatGPT got more score than Google Translate. Therefore, based on our analysis and the BLEU and METEOR scores it is concluded that ChatGPT performed better than Google Translate.

The text in Table 2 is an example of English to Urdu translation taken from the book *Genghis Khan: The Emperor of All Men* by Lamb [22]. We observed that neither translation is fully accurate, as both contain errors. They fail to fully convey the intended meaning of certain lines in this text. For instance, the translation of the sentence, “Children of this corner of the northern Gobi were not hardened to suffering; they were born to it,” is incorrect. The correct version is provided in the human translation provided by professional translator. Additionally, Google Translate was unable to translate some words into Urdu instead generated them as it is. For example, the meaning of the word Reed-bordered is present in Urdu language as سرکنڈوں سے گھری but Google Translate rendered the English word in Urdu script as ریٹہ بارڈر instead of translating it. However, ChatGPT accurately translated Reed-bordered as سرکنڈوں سے گھری and correctly captured the meaning of winged creatures as پرندے.

Table 1. Islamic text (Urdu to English translation).

Original Text		اسلام میں حقوق پر بڑا زور دیا گیا ہے۔ اور اسلامی ریاست شہریوں کے حقوق کی ذمہ دار قرار دی گئی ہے۔ اللہ تعالیٰ نے قرآن مجید میں اور حضرت محمد نے اپنی احادیث مبارکہ میں شہریوں کے حقوق کے حوالے سے بڑی واضح ہدایات فرمائی ہیں۔ انسانوں کو پابند کیا گیا ہے کہ وہ اسلام میں داخل ہو کر فرائض ادا کریں جو اللہ نے ان پر عائد کئے ہیں۔
ChatGPT Translation		In Islam, great emphasis has been placed on rights. The Islamic state is held responsible for ensuring the rights of its citizens. Allah Almighty has given clear guidelines in the Holy Quran, and Prophet Muhammad (peace be upon him) has provided explicit instructions in his noble hadiths regarding the rights of citizens. Humans have been bound to fulfill the duties imposed upon them by Allah after entering Islam.
ChatGPT Score	BLEU	0.5
	METEOR	0.216
Google Translation		Islam places great emphasis on rights. And the Islamic state is responsible for the rights of its citizens. Allah Almighty has given clear instructions regarding the rights of citizens in the Holy Quran and Hazrat Muhammad in his blessed hadiths. Humans are obligated to enter Islam and perform the duties that Allah has imposed on them.
Google Translate Score	BLEU	0.48
	METEOR	0.173
Human Translation		Rights have been greatly emphasized in Islam and Islamic state has been declared responsible for the protection of citizen's rights. Allah Almighty in the Holy Quran and Hazrat Muhammad (peace be upon him) in Ahadees have instructed much clearly regarding the rights of citizens. Men have been bound to pay duties bestowed by Almighty Allah after embracing Islam.

If we compare the BLEU and METEOR scores of ChatGPT and Google Translate presented in Table 2, we noticed that ChatGPT achieved higher scores than Google Translate. Therefore, based on our analysis and the BLEU and METEOR scores it is concluded that ChatGPT performed better than Google Translate.

3.1. Summary of Discussion

The results of the analysis can be summarized as follows:

1. Overall, ChatGPT performed better than Google Translate.
2. Out of 16 paragraphs (Annexure A), ChatGPT had higher BLEU and METEOR scores for 8 translations, Google Translate performed better for 5, and both had the same scores for

the remaining 3. Therefore, ChatGPT generally outperformed Google Translate.

3. ChatGPT's average BLEU score is 0.370, while Google Translate's is 0.353, as shown in Figure 4. Similarly, ChatGPT's average METEOR score is 0.322 and Google Translate's average METEOR score is 0.313, as presented in Figure 5.
4. It was observed that ChatGPT sometimes introduces extra content in its translations that are not present in the original text.
5. We noticed that ChatGPT improves its translation quality by learning from users' feedback; whereas Google Translate produces the same translation every time. We observed that AI-powered translators perform well when translating simple paragraphs, such as news articles. However, they struggle with complex texts, such as idioms.

Table 2. Text from novel (English to Urdu translation).

Original Text		“Life did not matter very much in the Gobi. Lofty plateaus, wind-swept, lying close to the clouds. Reed bordered lakes, visited by migratory winged creatures on their way to the northern tundras. Children of this corner of the northern Gobi were not hardened to suffering; they were born to it.”
ChatGPT Translation		گوبی میں زندگی کی زیادہ اہمیت نہیں تھی۔ بلند و بالا سطح مرتفع، ہوا سے جھلسے ہوئے، بادلوں کے قریب واقع۔ سرکنتوں سے گھری جھیلیں، جہاں ہجرت کرنے والے پرندے شمالی ٹنڈرا کی جانب جاتے ہوئے قیام کرتے تھے۔ شمالی گوبی کے اس گوشے کے بچے تکلیف کے عادی نہیں تھے؛ وہ تو اسی میں پیدا ہوئے تھے۔
ChatGPT Score	BLEU	0.288
	METEOR	0.272
Google Translation		گوبی میں زندگی بہت زیادہ اہمیت نہیں رکھتی تھی۔ اونچی سطح مرتفع، ہوا سے بھری ہوئی، بادلوں کے قریب پڑی ہے۔ ریڈ ہارٹر والی جھیلیں، جن کا دورہ شمالی ٹنڈراس کے راستے میں نقل مکانی کرنے والے پروں والے جانور کرتے ہیں۔ شمالی گوبی کے اس کونے کے بچوں کو تکالیف سے دوچار نہیں کیا گیا تھا۔ وہ اس سے پیدا ہوئے تھے۔
Google Translate Score	BLEU	0.270
	METEOR	0.177
Human Translation		گوبی میں زندگی کی کوئی اہمیت نہیں تھی۔ اونچے بلند ہموار ٹیلے، جن پر تیز ہواؤں کے جھکڑ جلتے، اور جن کی بلندی بادلوں کے قریب قریب پہنچتی۔ جھیلیں جن کے اطراف اونچی اونچی گھاس تھی۔ جن میں ہجرت کرنے والے پرندے شمالی ٹنڈراؤں کی طرف اڑتے ہوئے ان کر بسیرا لیتے۔ شمالی گوبی کے اس گوشے کی اولاد جو انسان تھے۔ انہیں تکلیفوں نے سخت جان نہیں بنایا تھا بلکہ سخت جانی ان کو ورثے میں ملی تھی۔

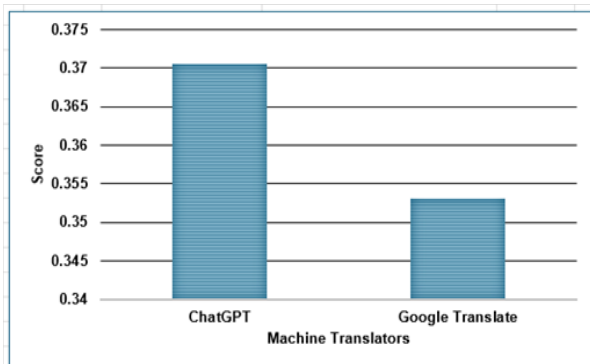


Fig. 4. The average BLEU score of ChatGPT and Google Translate.

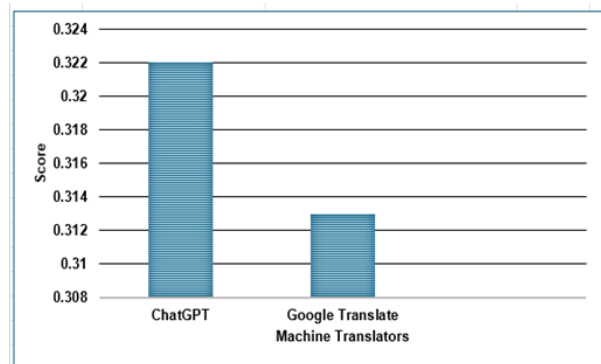


Fig. 5. The average METEOR score of ChatGPT and Google Translate.

4. CONCLUSIONS

Several online translation tools are available that offer reasonable accuracy, are free to use, and support a wide range of languages. Out of these, Google Translate remains the most widely utilized,

while OpenAI's ChatGPT has recently gained significant attention. These translation systems are capable of processing complex languages such as Urdu, Chinese, and Arabic. Machine translation tools can be particularly beneficial for multinational organizations, facilitating communication with

partners from diverse linguistic backgrounds. Additionally, they can also help students in translating and understanding content. We evaluated and compared the translations of 16 paragraphs, from English to Urdu and Urdu to English, generated by ChatGPT and Google Translate. Based on the evaluations in the preceding section, ChatGPT outperformed Google Translate, achieving higher METEOR and BLEU scores. ChatGPT's average BLEU score was 0.370, while Google Translate's was 0.353. Similarly, ChatGPT's average METEOR score was 0.322, compared to 0.313 for Google Translate. This study provides valuable insights for users seeking reliable online machine translation between English and Urdu and vice versa. To compare these translations, we used reference translations carefully crafted by experts in this domain. Mostly the machine-generated and human-generated translations convey a similar meaning to some extent, but they differ in terms of phrasing style and word selection, for example, in the translation of "Poetry" (Annexure A), the human translator chose the words "decree" and "ascertain", which preserve the depth and nuance of the original text. In contrast, the machine-generated translations use simple sentences to ensure general accuracy, but they often lack the depth and essence which we can feel in human translation. We observed that Google Translate was unable to translate some English words into Urdu and retained them as it is in Urdu script. For example, in the English to Urdu paragraph, the meaning of the word Reed-bordered in Urdu language is سرکنڈوں سے گھری but Google Translate rendered the English word in Urdu script as ریڈ بارڈر instead of translating it. In contrast, ChatGPT successfully translated most of the text. Overall, ChatGPT demonstrated higher translation accuracy compared to Google Translate for paragraph translation. However, there is always room for improvement, as some lines translated by ChatGPT were inaccurate and not giving the actual sense, as discussed in the previous section. Similarly, the translating idioms is a major challenge for these machine translators because they do not understand the cultural background or context in which they were expressed. One key feature of the ChatGPT is its ability to learn from user feedback and adjust its responses accordingly, which helps improve its translation accuracy. For example, for the translation of some text, we prompted ChatGPT with commands like "Translate again" or "Revise the translations" to enhance accuracy. After a few

attempts, it began generating correct translations. In the future, we plan to explore prompt engineering techniques to further improve translation accuracy.

5. ETHICAL STATEMENT

This study did not involve human participants or animals; therefore, ethical approval was not required.

6. FUNDING

No funding was received to conduct this research.

7. AUTHORSHIP CONTRIBUTION

All authors contributed equally to this research. They equally contributed to conceptualization, research methodology, data collection, implementation of translation experiments using Google Translate and ChatGPT-3.5, and the assessment of translation quality. The manuscript was primarily written by Aatka Faryal Riaz. The literature review was conducted by Rubeka Sehar. Validation of results, guidance, and final manuscript review were carried out by Rafay Azmat.

8. DECLARATION STATEMENT

We confirm that the results presented in this manuscript are original; these results are based on the study of Aatka Faryal Riaz, as part of M.Sc. Research Report, submitted to the International University of Applied Science, Bad Honnef, Germany; and it has not been published elsewhere nor is it under consideration for publication by another journal. All authors have approved the manuscript and agree that in case the article is accepted for publication, its copyright will be assigned to the *Pakistan Academy of Sciences*. Authors have obtained permission to reproduce, where needed, copyrighted material from other sources and ensured that no copyrights have been infringed upon.

9. REFERENCES

1. Ethnologue. *The Ethnologue 200* (2025). <https://www.ethnologue.com/insights/ethnologue200/>
2. International Center for Language Studies. *Most Spoken Languages in the World* (2024). <https://www.icls.edu/blog/most-spoken-languages-in-the-world/>
3. A.A. Malik and A. Habib. Urdu to English machine translation using bilingual evaluation understudy. *International Journal of Computer Applications*

- 82(7): 5-12 (2013). <https://research.ijcaonline.org/volume82/number7/pxc3891040.pdf>
4. M. Ghassemiazghandi. An evaluation of ChatGPT's translation accuracy using BLEU score. *Theory and Practice in Language Studies* 14(4): 985-994 (2024). <https://tpls.academypublication.com/index.php/tpls/article/view/7867/6371>
 5. S.S. Biswas. Potential use of ChatGPT in global warming. *Annals of Biomedical Engineering* 51(6): 1126-1127 (2023). <https://doi.org/10.1007/s10439-023-03171-8>
 6. S.S. Biswas. Role of ChatGPT in public health. *Annals of Biomedical Engineering* 51(5): 868-869 (2023). <https://link.springer.com/article/10.1007/s10439-023-03172-7>
 7. P.P. Ray. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 3: 121-154 (2023). <https://doi.org/10.1016/j.iotcps.2023.04.003>
 8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems* 30: (2017). <https://doi.org/10.48550/arXiv.1706.03762>
 9. P. Memon. Discover how ChatGPT is trained! (2023). <https://www.linkedin.com/pulse/discover-how-chatgpt-istrained-pradeep-memon>
 10. S.C. Siu. ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. *SSRN* (2023). <https://dx.doi.org/10.2139/ssrn.4448091>
 11. C. Boitet, H. Blanchon, M. Seligman, and V. Bellyneck. Evolution of MT with the Web. *Proceedings of the International Conference "Machine Translation 25 Years On", 21-22 November 2009, Cranfield, England* (2009). https://lig-membres.imag.fr/blanchon/Pdfs/MT25YO-Evolution-MT-web.pdf?utm_source=chatgpt.com
 12. I. Caswell. Google Translate learns 24 new languages. *Online google blog* (2022). <https://blog.google/products/translate/24-new-languages/>
 13. Google Translate. *Google Translate Wikipedia* (2025). https://en.wikipedia.org/wiki/Google_Translate
 14. P. Chang, P.J. Chen, and L.L. Lai. Recursive editing with Google Translate: the impact on writing and error correction. *Computer Assisted Language Learning* 37(7): 2116-2141 (2024). <https://doi.org/10.1080/09588221.2022.2147192>
 15. S.C. Tsai. Using Google Translate in EFL drafts: A preliminary investigation. *Computer Assisted Language Learning* 32(5-6): 510-526 (2019). <https://doi.org/10.1080/09588221.2018.1527361>
 16. K.M. King. Can Google Translate be taught to translate literature? A case for humanists to collaborate in the future of machine translation. *Translation Review* 105(1): 76-92 (2019). <https://doi.org/10.1080/07374836.2019.1673268>
 17. Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M.N. Yonghui, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv: 1609.08144* (2016). <https://doi.org/10.48550/arXiv.1609.08144>
 18. K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA* (2002). <https://aclanthology.org/P02-1040.pdf>
 19. M.A. Ayesha, S. Noor, M. Ramzan, H.U. Khan, and M. Shoaib. Evaluating Urdu to Arabic machine translation tools. *International Journal of Advanced Computer Science and Applications* 8(10): 90-96 (2017). https://thesai.org/Downloads/Volume8No10/Paper_12-Evaluating_Urdu_to_Arabic_Machine_Translation_Tools.pdf
 20. S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, USA* (2005). https://aclanthology.org/W05-0909.pdf?utm_source=chatgpt.com
 21. S. Khan, N. Aftab, A. Ali, and R. Mohammad. Mutalia Pakistan, Class 12. *Punjab Textbook Board, Lahore, Pakistan* (2005). <https://www.taleem360.com/12th-class-pak-study-new-punjab-text-book-2022-zqx>
 22. H. Lamb (Ed.). *Genghis Khan, the Emperor of All Men*. *Doubleday, New York, USA* (1927). https://archive.org/details/genghiskhanemperor0000lamb_f6b1/page/n13/mode/2up