Journal of Translation and Language Studies
https://jtls.sabapub.com

2023 Volume 4, Issue 1: 1 – 23
DOI : https://doi.org/10.48185/jtls.v4i2.762

Exploring the Effectiveness of GPT-3 in Translating Specialized Religious Text from Arabic to English: A Comparative Study with Human Translation

Maysaa Banat[1*], Yasmine Abu Adla[2]

[1]Rafik Hariri University, banatms@rhu.edu.lb
[2] American University of Beirut, yaa41@mail.aub.edu

Received: 13.05.2023 • Accepted: 20.06.2023 • Published: 14.07.2023 • Final Version: 17.07.2023

Abstract: In recent years, Natural Language Processing (NLP) models such as Generative Pretrained Transformer 3 (GPT-3) have shown remarkable improvements in various languagerelated tasks, including machine translation. However, most studies that have evaluated the performance of NLP models in translation tasks have focused on generalpurpose text, leaving the evaluation of their effectiveness in handling specialized text to be relatively unexplored. Therefore, this study aimed to evaluate the effectiveness of GPT-3 in translating specialized Arabic text to English and compare its performance to human translation.

To achieve this goal, the study selected ten chapters from a specialized book written in Arabic, covering topics in specialized religious context. The chapters were translated by a professional human translator and by GPT-3 using its translation Application Programming Interface. The translation performance of GPT-3 to was compared to human translation using qualitative measures, specifically the Direct Assessment method. Additionally, the translations were evaluated using two different evaluation metrics, Bidirectional Encoder Representations from Transformers (BERT) score and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric, which measure the similarity between the translated text and the reference text.

The qualitative results show that GPT produced generally understandable translations but failed to capture nuances and cultural context. On the other hand, the quantitative results of the study showed that GPT-3 was able to achieve a relatively high level of accuracy in translating specialized religious text, with comparable scores to human translations in some cases. Specifically, the BERT score of GPT-3 translations was 0.83. The study also found that the Rouge score failed to fully reflect the capabilities of GPT-3 in translating specialized text.

Overall, the findings of this study suggest that GPT-3 has promising potential as a translation tool for specialized religious text, but further research is needed to improve its capabilities and address its limitations.

Keywords: GPT3, Machine Translation, Arabic to English Translation, BERT Score, ROUGE Score, Natural Language Processing

* Corresponding Author: banatms@rhu.edu.lb

1. Introduction

NLP is a field of study that focuses on developing computer algorithms capable of processing and understanding human language. NLP has seen significant progress in recent years with the development of advanced deep learning models such as GPT-3, which uses a large language model (LLM) to generate human-like text. GPT-3, a large language model trained by OpenAI, is a prime example of the advancements made in NLP and deep learning. Language models like ChatGPT use vast amounts of data to learn and understand patterns in human language, making them capable of generating text that is almost indistinguishable from that written by humans. These models have shown great promise in a wide range of applications, from chatbots that can understand and respond to natural language queries to language translation and even writing automated news articles [1].

Moreover, the development of NLP and language models such as ChatGPT has opened up a whole new world of possibilities for businesses, governments, and organizations worldwide. These advancements enable machines to process and analyze natural language data in ways that were once impossible [2]. With the ability to process vast amounts of textual data, machines can now provide insights into customer sentiment, automate customer service, and even detect fraud. As such, NLP and ChatGPT have become integral tools for organizations seeking to harness the power of natural language data and transform their operations.

The translation of specialized text poses a significant challenge for machine translation systems, as it requires a deep understanding of the domain-specific terminology and the ability to accurately convey the meaning of the text. This has led to a growing need to develop machine translation systems that can handle specialized text accurately. However, evaluating the accuracy of machine translation systems is a crucial step in the development of effective translation tools. Various evaluation metrics, such as Bilingual Evaluation Understudy (BLEU), Metric For Evaluation Of Translation With Explicit Ordering (METEOR), Rouge, and Bert score have been proposed to assess the accuracy of machine translation. These metrics are used to compare the machine-generated translation with a reference translation, which is typically a human-generated translation [3,4].

The translation of religious text requires a deep understanding of the language and cultural context, as well as a familiarity with religious terminology and concepts. As such, it presents a unique challenge for machine translation systems such as GPT-3. The accuracy and fluency of machine translation can be affected by the complexity and specificity of religious terminology and expressions, as well as the need for a coherent and cohesive translation that conveys the intended meaning of the original text.

Therefore, evaluating the effectiveness of GPT-3 in translating specialized religious text from Arabic to English is an important step in understanding the capabilities and limitations of machine translation in this domain. This study aims to evaluate the translation quality of the GPT-3 language model and compare it to human translation using both quantitative and qualitative measures. The study seeks to provide a comprehensive understanding of the strengths and weaknesses of GPT-3's translations and to identify areas where human translators still outperform machine translation systems. The findings of this study will contribute to our understanding of the potential uses and limitations of machine translation in the translation of religious texts, and inform the development of machine translation systems that can produce accurate and effective translations in this specialized domain.

2. Literature Review

The following literature review provides an overview of the existing research and current knowledge on the topic of evaluating machine translation. It highlights the key findings, trends, and gaps in the literature, as well as the most prominent theories and approaches that have been used to study this area.

2.1. Machine Learning and its Emergence in Translation

Machine learning has become increasingly popular in translation and has been widely used in NLP. Traditional rule-based machine translation systems relied on human-crafted rules to translate text. However, these systems were limited in their ability to handle language variation and context. Machine learning-based translation, on the other hand, is capable of learning patterns and rules from large datasets. This has led to significant improvements in translation quality.

Neural Machine Translation

. Neural Machine Translation (NMT) is a type of machine learning-based approach to translation that has gained popularity in recent years. NMT uses artificial neural networks to learn the mapping between the source language and the target language. The neural network consists of multiple layers of interconnected nodes, each layer processing the input from the previous layer to gradually transform the input into the desired output [2].

Advantages of NMT

. One of the key advantages of NMT is that it can learn to handle long-range dependencies between words in a sentence. Unlike previous rule-based and statistical machine translation approaches that often struggled with long and complex sentences, NMT models can also learn to handle sentence-level and context-level translations. This means that they can better capture the nuances and idiomatic expressions of human language. NMT has become the state-of-the-art approach for many translation tasks due to its impressive performance on various language pairs. The neural network architecture of NMT models can be customized for specific languages, domains, and tasks, making it a flexible approach for translation. In addition, such models can be trained on large parallel corpora, which allows them to learn from a vast amount of data, leading to improved translation quality [5].

Limitations of NMT

. Despite its many advantages, NMT also has some limitations. Training an NMT model can be computationally expensive, and the quality of the translation output can be heavily dependent on the quality and quantity of the training data. In addition, such models may struggle with rare or unseen words and may produce unnatural or incorrect translations in some cases [6].

Pre-Trained Language Models

. In recent years, the development of large pre-trained language models, such as GPT-3, has revolutionized the field of NLP, including machine translation. These models have been trained on massive amounts of text data, enabling them to generate human-like text with remarkable accuracy and fluency.

Fine-Tuning GPT-3

. One of the main advantages of pre-trained language models is that they can be fine-tuned for a specific task, such as translation, with only a small amount of task-specific training data. This is known as transfer learning, and it has enabled researchers and practitioners to develop high-quality translation models with relatively little effort and time. GPT-3 has demonstrated impressive performance on a range of language tasks, including translation, and has become a popular choice for many researchers and practitioners in the field of NLP. The model uses a large transformer-based architecture that allows it to process and generate text with a high level of accuracy and fluency. One of the key features of GPT-3 is its ability to generate text that is contextually relevant and coherent, even in the absence of explicit rules or prompts. This makes it a powerful tool for tasks such as machine translation, where the context and meaning of the source text must be accurately captured and conveyed in the target language [7].

Limitations of Pre-Trained Language Models.

While pre-trained language models like GPT-3 have shown remarkable performance on a range of language tasks, they also have some limitations. These models can be computationally expensive to train and may require specialized hardware to achieve the best performance. In addition, the quality of the translation output can still be heavily dependent on the quality and quantity of the training data.

Machine learning has been widely used in NLP and is becoming increasingly popular in translation. Traditional rule-based machine translation systems relied on human-crafted rules to translate text, which made them limited in their ability to handle language variation and context. Machine learning-based translation, on the other hand, is capable of learning patterns and rules from large datasets, which has led to significant improvements in translation quality.

Neural machine translation (NMT) is a type of machine learning-based approach to translation that has gained popularity in recent years due to its impressive performance on various language pairs. Such artificial neural networks are used to learn the mapping between the source language and the target language. The neural network consists of multiple layers of interconnected nodes, each layer processing the input from the previous layer to gradually transform the input into the desired output.

One of the key advantages of NMT is that it can learn to handle long-range dependencies between words in a sentence, unlike the previous rule-based and statistical machine translation approaches that often struggled with long and complex sentences.These models can also learn to handle sentence-level and context-level translations, which means that they can better capture the nuances and idiomatic expressions of human language.

NMT has become the state-of-the-art approach for many translation tasks due to its impressive performance on various language pairs. The neural network architecture of NMT models can be customized for specific languages, domains, and tasks, making it a flexible approach for translation. In addition, such models can be trained on large parallel corpora, which allows them to learn from a vast amount of data, leading to improved translation quality.

Despite its many advantages, NMT also has some limitations. Training an NMT model can be computationally expensive, and the quality of the translation output can be heavily dependent on the quality and quantity of the training data. In addition, such models may struggle with rare or unseen words and may produce unnatural or incorrect translations in some cases.

Overall, NMT has shown impressive results in machine translation, and its flexibility and ability to handle long-range dependencies and context make it a promising approach for future research in machine translation [2].

In recent years, the development of large pre-trained language models, such as GPT-3, has revolutionized the field of NLP, including machine translation. These models have been trained on massive amounts of text data, enabling them to generate human-like text with remarkable accuracy and fluency.

One of the main advantages of pre-trained language models is that they can be finetuned for a specific task, such as translation, with only a small amount of task-specific training data. This is known as transfer learning, and it has enabled researchers and practitioners to develop high-quality translation models with relatively little effort and time.

GPT-3 has demonstrated impressive performance on a range of language tasks, including translation, and has become a popular choice for many researchers and practitioners in the field of NLP. The model uses a large transformer-based architecture that allows it to process and generate text with a high level of accuracy and fluency. One of the key features of GPT-3 is its ability to generate text that is contextually relevant and coherent, even in the absence of explicit rules or prompts. This makes it a powerful tool for tasks such as machine translation, where the context and meaning of the source text must be accurately captured and conveyed in the target language.

While pre-trained language models like GPT-3 have shown remarkable performance on a range of language tasks, they also have some limitations. These models can be computationally expensive to train and may require specialized hardware to achieve the best performance. In addition, the quality of the translation output can still be heavily dependent on the quality and quantity of the training data, as well as the choice of fine-tuning strategy.

Overall, the development of pre-trained language models like GPT-3 has had a significant impact on the field of NLP, including machine translation. These models have enabled researchers and practitioners to develop high-quality translation models with relatively little effort and time, and they continue to drive innovation in the field. [1].

Despite the promising results of machine learning-based translation, there are still challenges that need to be addressed. One major challenge is the lack of large high-quality parallel datasets for many language pairs and specialized domains, which can limit the effectiveness of machine learning-based translation. Furthermore, machine learning-based translation systems are often seen as black boxes, making it difficult to understand how they make translation decisions, which can be a concern in some applications.

In conclusion, machine learning-based translation has emerged as a promising approach for addressing the limitations of traditional rule-based machine translation. The development of NMT and pre-trained language models has shown significant improvements in translation quality. However, there are still challenges that need to be addressed, such as the availability of large high-quality parallel datasets and the interpretability of machine learning-based translation systems.

ChatGPT: A Powerful Language Model for Translation and Conversation.

ChatGPT is a large language model developed by OpenAI based on the GPT-3.5 architecture. It is designed to be able to generate natural and coherent responses to a wide range of inputs, including text prompts, questions, and conversation [8]. GPT-3 is based on a transformer architecture, which is a type of neural network that is particularly well-suited to language processing tasks. The

transformer architecture was first introduced in a paper by Vaswani et al. in 2017, and has since become a popular choice for language models [9].

The GPT-3 model has a massive number of parameters, with the largest version of the model having over 175 billion parameters. This makes it one of the largest and most powerful language models in existence. The model is trained on a diverse range of text data, including books, articles, and websites, in order to learn patterns and relationships in language.

At a high level, as seen in Figure 1, the GPT-3 architecture consists of a series of transformer blocks that are connected in a feedforward network. Each transformer block has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feedforward network. The self-attention mechanism allows the model to attend to different parts of the input text when making predictions, while the feedforward network applies a non-linear transformation to the input features.

The GPT-3 model uses a technique called "unsupervised pre-training" to learn the patterns and relationships in language. During pre-training, the model is trained on a large corpus of text data using a language modeling objective. The objective is to predict the next word in a sequence given the previous words in the sequence. By training the model to predict the next word in a sequence, the model learns to understand the patterns and relationships in language.

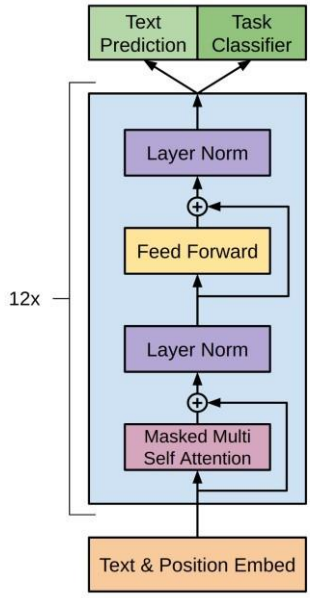


Figure 1. GPT3 Architecture. Adopted from [10]

To use ChatGPT for translation, the input text is first tokenized into a sequence of tokens, and each token is assigned an embedding vector that represents its meaning. These embeddings are then fed into the model's encoder, which generates a hidden representation of the text by processing the embeddings through multiple layers of self-attention and feedforward networks. Once the input text has been encoded, the model's decoder is used to generate the output text in the target language. The decoder works by generating one token at a time, based on the patterns learned during training on a large corpus of text data. The decoder uses the hidden representation generated by the encoder, as well as a special token representing the target language, to generate each token in the output sequence. During training, ChatGPT is optimized to minimize the difference between the generated output and the actual target text. This is done by adjusting the weights of the model's parameters

using backpropagation and gradient descent. The model is typically trained on a large corpus of text data, which allows it to learn the patterns and structures of the language(s) it is trained on [8].

While ChatGPT may not be as accurate as dedicated machine translation models for specific language pairs, it is often capable of generating more fluent and natural-sounding translations due to its ability to generate text based on context and semantics. This is because ChatGPT is able to leverage its large training corpus to capture the nuances of the language(s) it is trained on, and can generate translations that take into account the broader context of the input text.

Overall, ChatGPT represents a significant advance in natural language processing technology, offering a powerful tool for generating natural and coherent text in a variety of contexts, including translation.

Advancements in Machine Translation Evaluation

In a study entitled " Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine" [11], the researchers reported an initial evaluation of ChatGPT's machine translation capabilities in terms of translation prompts, multilingual translation, and translation robustness. The study assessed the effectiveness of multilingual translation systems on 7 test sets that contain a total of 1012 sentences translated into 101 languages. The test sets included the WMT19 (Workshop on Machine Translation) Biomedical Translation Task (Bio) and the WMT20 Robustness Task (Rob2 and Rob3). Due to time limitations, the authors randomly selected 50 sentences from each set for evaluation. The primary evaluation metric used is the BLEU score, but ChrF++ (Character n-gram F-score) and TER (Translation Error Rate) were also occasionally reported; these metrics were supported by SacreBLEU. Additionally, the authors evaluated four commonly used languages (German, English, Romanian, and Chinese) by testing translation performance in 12 different directions and comparing BLEU scores to those of Google Translate. While ChatGPT's translation robustness is not as good as commercial systems for biomedical abstracts or Reddit comments, it performed well on spoken language. However, with the release of the GPT-4 engine, ChatGPT's translation performance has been greatly improved, and it is now comparable to commercial translation products even for distant languages.

In another study titled "Document-Level Machine Translation with Large Language Models" [12], the researchers examined the ability of large language models, such as ChatGPT, to generate coherent and fluent translations for document-level machine translation. The study focused on three aspects: the effects of discourse-aware prompts, the comparison of translation models, and the analysis of discourse modeling abilities. They evaluated different approaches and systems using both sentence- and document-level evaluation metrics. Regarding sentence-level metrics, they employed the commonly-used sacreBLEU and TER. In addition, they utilized COMET, which leverages pretrained language models to achieve high correlations with human quality judgments. For document-level metrics, they reported document-level sacreBLEU, which was computed by matching ngrams in the whole document. Through evaluation of a number of benchmarks, the study found that ChatGPT outperforms commercial MT systems in terms of human evaluation, and GPT-4 demonstrated a strong ability to explain discourse knowledge. Despite some challenges in selecting correct translation candidates in contrastive testing, ChatGPT and GPT-4 have demonstrated superior performance and potential to become a new paradigm for document-level translation. The study emphasized the challenges and opportunities of discourse modeling for LLMs, which could inspire future research in this area.

Taking recent advancements in context-aware neural machine translation into consideration, the study titled "Do Online Machine Translation Systems Care for Context? What About a GPT Model?" [13] sheds light on the challenges associated with evaluating document-level machine translation [13].The study investigated how well online machine translation systems deal with six context-related issues: lexical ambiguity, grammatical gender, grammatical number, reference, ellipsis, and terminology, when a larger context span containing the solution for those issues was given as input. The study compared the results of online Machine Translation systems to the translation outputs from ChatGPT and found that although the change of punctuation in the input yielded great variability in the output translations, the context position did not seem to have a significant impact. Additionally, the study revealed that the GPT model outperformed the NMT systems, but its performance for Irish was poor.

In their research paper "Empirical Results and Analysis of Multilingual Machine Translation with Large Language Models" [14], the authors delved into the potential of LLMs in multilingual machine translation and examined the factors that influence their translation performance. The study evaluated several popular LLMs, including XGLM, OPT, BLOOMZ, and ChatGPT, on 102 languages. The results demonstrated that, despite the current best-performed LLM being ChatGPT, it still laged behind the supervised baseline "No Language Left Behind" model in 83.33% of translation directions. The study identified new working patterns of LLMs in MMT, including the surprising ability to ignore prompt semantics when given in-context exemplars and the effectiveness of cross-lingual exemplars for low-resource translation. The study also highlighted the potential risk of using public datasets for evaluation, as observed in the overestimated performance of BLOOMZ on the Flores-101 dataset.

In their research titled "A Comprehensive Evaluation of Generative Pre-trained Transformer (GPT) Models for Machine Translation" [1], the authors conducted a detailed examination of the efficacy of GPT models in machine translation. The evaluation covered several aspects, including the quality of different GPT models compared to state-of-the-art research and commercial systems, the effect of prompting strategies, robustness towards domain shifts, and document-level translation. The experiment involved eighteen different translation directions, high and low resource languages, and non-English-centric translations, and the performance of three GPT models, namely ChatGPT, GPT3.5 (text-davinci003), and text-davinci-002, was evaluated. The results indicated that GPT models achieve competitive translation quality for high resource languages, but their capabilities for low resource languages are limited. The study also demonstrated that hybrid approaches that combine GPT models with other translation systems can further enhance translation quality. Comprehensive analysis and human evaluation were conducted to gain a better understanding of the characteristics of GPT translations. The study provided valuable insights for researchers and practitioners in the field to understand the potential and limitations of GPT models for translation.

In their work, "Evaluating the Quality of Machine Translation: A Comprehensive Review of Automated and Human Metrics" [15], the authors provided an extensive examination of the current automated, semi-automated, and human metrics utilized to assess the quality of machine translation output. Divided into three parts, the first section covered reference translation-based metrics, confidence or quality estimation metrics, and diagnostic evaluation based on linguistic checkpoints. The second part discussed human metrics, which are classified based on whether human judges express a subjective evaluation judgment or not. The article provided details on tasks such as fluency and adequacy annotation, ranking, direct assessment (DA), error classification, and post-editing. The final section focused on the specific challenges posed by NMT and suggests ways to evaluate NMT

based on the latest research. The article emphasized the importance of human evaluation metrics and their relevance to interdisciplinary research groups that evaluate MT systems.

3. Methodology

The methodology section outlines the procedures and techniques utilized in this research study to investigate the effectiveness of the natural language processing model GPT-3 in translating specialized Arabic text to English. This section will detail the study design, data collection methods, and the analytical tools employed to evaluate the accuracy of GPT-3 translations in comparison to human translations. By providing a clear and comprehensive description of the research methodology, this section aims to ensure the transparency and replicability of the study's findings.

Qualitative measures were used to evaluate the translation quality of GPT-3. Specifically, we compared the translations produced by GPT-3 to those created by human translators using direct assessment. This approach involved having human judges rate the quality of translations based on factors such as fluency, accuracy, and naturalness. By using this approach, we were able to gain a more nuanced understanding of the strengths and weaknesses of GPT-3's translations and identify areas where human translators still outperformed machine translation systems.

In addition to using qualitative metrics, this study aimed to evaluate the translation abilities of the GPT-3 language model on specialized Arabic text. As seen in Figure 2, the methodology involved translating ten chapters from Arabic to English using a specialized human translator, and then translating the same chapters using GPT-3. To assess the performance of GPT-3, the generated translations were compared to the human translations using BERT and ROUGE scores. These scores were analyzed using a paired t-test to determine the statistical significance of the difference between the two translation methods. The results were used to quantify GPT-3's translation abilities on specialized text, and to determine whether it is more robust than other machine translation models. Overall, this methodology allowed for a comprehensive evaluation of GPT-3's performance on specialized text, providing valuable insights into the capabilities and limitations of this cutting edge language model for translation tasks.

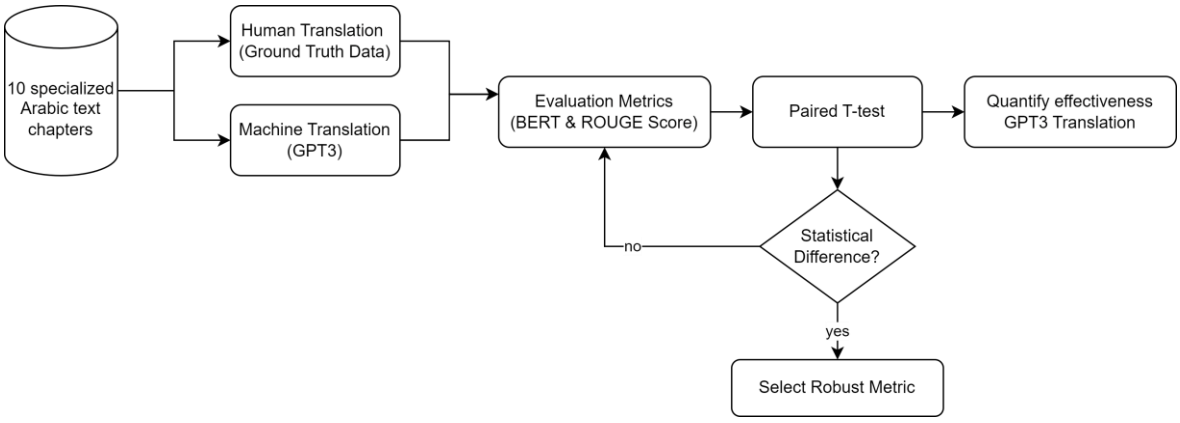


Figure 2. Flowchart of Methodology.

Dataset Description

. The dataset serves as a valuable resource for evaluating the effectiveness of natural language processing models, particularly in handling specialized religious text. In this study, a dataset

comprising a religious book of 239 pages has been utilized, which includes a total of 30 chapters along with an introductory chapter and a closure. Each chapter is comprised of approximately 7 pages and contains around 5 paragraphs per page, resulting in an average of 1500 words per chapter. The book incorporates a diverse range of religious texts such as Hadith and Quranic verses, and each chapter comprises Hadith verification/sourcing [16].

In this study, the focus was specifically on ten chapters out of the total thirty chapters. For the human translated chapters, each one is approximately 9 pages long, with around 5 paragraphs per page, resulting in approximately 1800 words per chapter. The translated chapters include hadith, Quranic verses, and hadith verification/sourcing.

As for the GPT-3 translated chapters, each chapter is approximately 8 pages long, with around 5 paragraphs per page, resulting in approximately 1700 words per chapter. The GPT-3 translated chapters include hadith, Quranic verses, and hadith verification/sourcing.

Analysis of Translation Quality: Human vs. Machine Translation

. In this study, we conducted an evaluation of the GPT translation of an Arabic text by comparing it to a human translation. The evaluation process was based on various criteria, including accuracy, fluency, cohesion and coherence, style and register, and word choice, with the human translation serving as the reference. Accuracy refers to the degree to which the translation conveys the intended meaning of the original text. Fluency measures how natural the translation sounds in the target language. Cohesion and coherence assess how well the translation is organized and connected. Style and register examine whether the translation is suitable for the intended audience and purpose. Lastly, word choice considers whether the translation uses the most appropriate words and expressions to convey the meaning of the original text accurately [15]. By conducting an evaluation of the GPT translation using these criteria, we aimed to gain insights into the strengths and weaknesses of machine translation and identify areas for improvement.

Measuring Translation Quality: Human vs. Machine Translation

.

To evaluate the quality of the machine translation and human translation, we used BERT Score and ROUGE, which are both metrics that measure the similarity between two sets of text. BERT Score uses contextual embeddings obtained from the BERT language model, while ROUGE measures the overlap of the longest common subsequence (LCS) between the machine translation output and the reference translations.

For the evaluation process, we compared the machine translation output to the reference translations (i.e., the original human translations). The reference translations were used as the gold standard for the evaluation. We used both BERT Score and ROUGE to compare the similarity of the machine translation output with the reference translations. We took the mean of the BERT Score and ROUGE for the sentences of each chapter and then computed the scores for the entire chapter. The resulting scores for each chapter were then compared to the gold standard.

After calculating the BERT Score and ROUGE-L scores for each machine translation output, we compared the scores to the scores obtained from the reference translations. We used statistical methods to test whether the differences between the scores were significant or not.

Overall, by using BERT Score and ROUGE-L, we were able to compare the quality of machine translation output with the quality of human translation, and to identify the strengths and weaknesses of each approach.

BERT Score

. The BERT Score metric measures the similarity between two sets of text based on contextual embeddings. Specifically, it uses a pre-trained BERT model to generate embeddings for each sentence in the machine translation output and the reference translations [3].

The embeddings are high-dimensional vectors that represent the contextual meaning of the sentence. They are generated by feeding the sentence through the BERT model, which uses self-attention to capture the relationships between the words in the sentence.

As seen in Figure 3, once the embeddings are generated for both sets of text, BERT Score calculates the cosine similarity between the embeddings of each sentence in the machine translation output and the reference translations. The cosine similarity measures the cosine of the angle between the two vectors, which indicates the similarity between the two sentences.

To weigh the similarity measures, BERT Score uses idf (inverse document frequency) weights, which assign lower weights to common words and higher weights to rare words. This helps to capture the importance of words that are more specific to the text being compared.

The sentence-level BERT Score is calculated by taking the average of the cosine similarities across all sentences in the machine translation output and the reference translations. The final score is obtained by averaging the sentence-level scores.

By using BERT Score, we were able to capture the semantic similarity between the machine translation output and the reference translations, which is important for evaluating the quality of machine translation. Compared to traditional metrics like BLEU, which only consider the overlap of n-grams between the two sets of text, BERT Score provides a more nuanced and accurate measure of the similarity between the machine translation output and the reference translations.

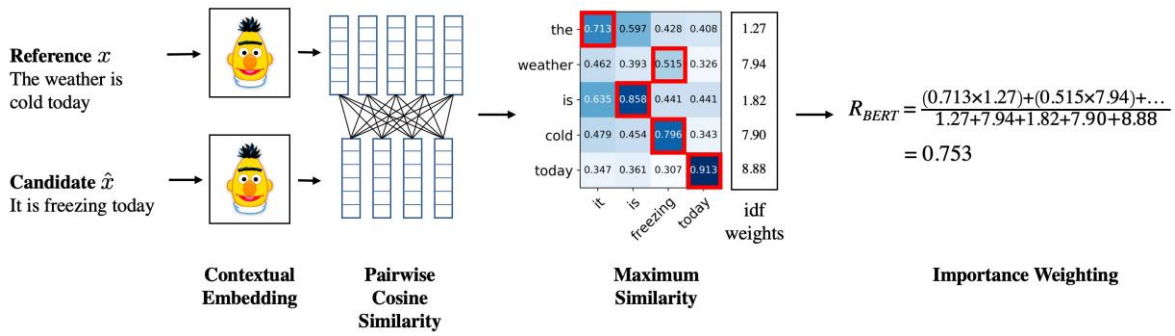


Figure 3. Illustration of the computation of the recall metric RBERT. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We weigh the similarity measures using idf weights. Adopted from [3]

ROUGE Score

. ROUGE is a set of metrics used to evaluate the quality of text summarization and machine translation output. ROUGE is based on the idea of comparing a generated summary or translation to a set of reference summaries or translations, and measuring the overlap between them. Specifically, the ROUGE score is calculated as the sum of the recalls of ngrams that appear in both the machine-generated summary and the reference summaries, where recall of n-grams is the number of n-grams in the reference summaries that are also present in the machine-generated summary, divided by the total number of n-grams in the reference summaries [4]. The ROUGE formula is shown in Equation 1.

$$ROUGE_L = \frac{\sum_{r \in \text{Reference Summaries}} \sum_{gram \in LCS(r,\hat{r})} \text{countmatch}(gram)}{\sum r \in \text{Reference Summaries} \sum_{gram \in r} \text{count}(gram)} \quad (1)$$

Where:

Reference Summaries refers to the set of reference summaries

$r\hat{}$ refers to the machine-generated summary

LCS($r,r\hat{}$) refers to the LCS between the reference summary r and the machine-generated summary $r\hat{}$

count($gram$) refers to the total number of occurrences of the n-gram $gram$ in the reference summaries

$\text{count}_{\text{match}}(gram)$ refers to the total number of occurrences of the n-gram $gram$ in both the machine-generated summary and the reference summaries

There are different variations of the ROUGE metric, with different ways of measuring overlap. In this study, ROUGE-L was used in this study, which is one popular variant that measures the LCS between the machine translation output and the reference translations. The LCS is the longest sequence of words that occurs in both the machine translation output and the reference translations, without necessarily being contiguous.

ROUGE-L calculates the F1-score of the LCS between the machine translation output and the reference translations. The F1-score is the harmonic mean of precision and recall, where precision is the number of words in the LCS divided by the number of words in the machine translation output, and recall is the number of words in the LCS divided by the number of words in the reference translations.

The ROUGE-L score ranges from 0 to 1, with 1 indicating perfect similarity. A ROUGEL score of 0 means that there is no overlap between the machine translation output and the reference translations.

Paired T-test

. To compare the performance of BERT Score and ROUGE-L, we used a paired t-test. For each chapter, we computed the BERT Score and ROUGE-L scores for all the translations and calculated the mean scores. Then, we performed a paired t-test on the mean scores to determine if there was a statistically significant difference between the two metrics.

The paired t-test is a statistical test that compares the means of two related groups. In our case, the two related groups are the BERT Score and ROUGE-L scores for each chapter. The paired t-test

takes into account the correlation between the two groups and allows us to determine if the difference between the means is statistically significant or due to chance [17].

If the p-value of the paired t-test is less than a predefined threshold (e.g., 0.05), we reject the null hypothesis that there is no significant difference between the means of the two metrics and conclude that one metric performs better than the other. By using the paired t-test, we were able to objectively compare the performance of BERT Score and ROUGE-L and determine which metric was more effective in evaluating the quality of our machine translations.

4. Results and Discussion

In this section, we will also conduct a qualitative analysis by closely examining sample outputs from both models to gain a deeper understanding of their strengths and weaknesses. Then, we will delve into the differences between ROUGE and BERT through the use of various visualizations and qualitative analysis. Specifically, we will employ bar graphs and box plots to visualize the quantitative differences between the two models. Through this multifaceted approach, we hope to provide a comprehensive and nuanced comparison between ROUGE and BERT.

Analysis of Translation Quality: Qualitative Assessment

. A qualitative analysis was conducted by closely examining sample outputs from both the human translator and GPT-3 models. This was done to gain a deeper understanding of the strengths and weaknesses of each approach.

Original Arabic Text:

المحبة، هي موجة الأثير التي تحمل رسائل المحبين، هي من أسرار قوة الروح، فمن أحب الله تعالى وأخلص له في النية والقول والعمل، سرت روحه بين أحبابها، فغذت أرواح الطالبين، ورفعت همة السالكين، ووسعت عقل المجالسين،وأراحت قلوب المحبين، وجذبت أرواح التائهين.

Human Translation

. . Love, the wave of ether that carries the messages of lovers, is one of the secrets of the power of the soul. Indeed, he who loves Allah Almighty and sincerely in intention, will be blessed with an empowered soul that will roam around the globe nurturing the souls of lovers, motivating them, and preaching them to the right track.

GPT-3 Translation

. Love is the etheric wave that carries messages of the lovers. It is one of the secrets of spiritual power. Whoever loves Allah and dedicates themselves to Him in intention, speech, and action, their soul will travel among its companions. It will nourish the souls of the seekers, elevate the aspirations of the seekers, expand the minds of the gatherings, ease the hearts of the lovers, and attract the souls of the lost.

Direct Assessment:

Accuracy:. both translations convey the main idea of the text, which is about the power of love and its ability to nurture souls. However, there are some differences in word choice and phrasing that affect the accuracy of the translations. For example, in the human translation, in the GPT-3 translation, the phrase "dedicate themselves to Him" is not accurate, as the original text refers to the love of Allah, not dedication.

Fluency. : both translations are grammatically correct and easy to read. However, the human translation is more natural and flows better, as it uses more idiomatic expressions and phrasing. For

example, the phrase "preaching them to the right track" in the human translation is more natural than "elevate the aspirations of the seekers" in the GPT-3 translation.

Cohesion and coherence. : both translations are coherent and well-structured. However, the human translation is more cohesive, as it uses more connectives and transitional phrases that help to link the different ideas and sentences together. For example, the phrase "Indeed, he who loves Allah Almighty" in the human translation provides a logical connection to the previous sentence, while in the GPT-3 translation, this link is not explicitly stated.

Style and register. : both translations use formal and religious language that is appropriate for the topic. However, the human translation is more stylistically varied, as it uses different types of sentence structures and rhetorical devices such as repetition and parallelism. For example, the repetition of the phrase "nurturing the souls of lovers, motivating them, and preaching them to the right track" in the human translation creates a sense of rhythm and emphasis that is not present in the GPT-3 translation.

Word Choice. : both translations use some accurate and appropriate vocabulary to convey the meaning of the original text. However, there are some differences in word choice that affect the accuracy and fluency of the translations. For example, in the human translation, the phrase "he who loves Allah Almighty" is more accurate than "whoever loves Allah and dedicates themselves to Him" in the GPT-translation. Additionally, the phrase "nurturing the souls of lovers" in the human translation is more accurate than "nourish the souls of the seekers" in the GPT-3 translation.

Original Arabic **Text:**

. الروح المطلقة هي روح كروح النبي ﷺ، حباه الله بها، لأن عبده ﷺ لم يترك لنفسه حظا، وجعل كلَّه لله تعالى، علماً وعملاً ومعاملةً وعبادة. فاطلق الله تعالى له روحه فجعلت تجوب الأكوان وترى بديعَ الخلق وعظيمَه، وتحوم فوق الأرض فترى وتسمع.

Human Translation

. An absolute soul resembles the prophetic soul; a gift that has been bestowed upon Allah’s messenger, peace be upon him, for being a true servant of Him. He, may Allah’s blessings be upon him, devoted his worldly and divine matters genuinely to Allah, Almighty, who released His prophet’s soul free of boundaries, and set it to roam the universe, explore, and hover over the earth to witness Allah’s grand creation.

GPT-3 Translation

. The absolute soul is like the soul of the Prophet, which Allah has blessed him with, because the Prophet (peace be upon him) did not leave anything for himself and dedicated everything to Allah, knowledge, actions, dealings and worship. So Allah granted him his soul and made it roam the universes, seeing the wonders of creation and its greatness, and hovering over the earth to see and hear.

Direct Assessment:

Accuracy. : the human translation is more faithful to the original Arabic text. For example, in the Arabic text, the phrase "الروح المطلقة" (al-ruh al-mutlaqa) refers to the concept of an absolute soul,

which is translated correctly by the human translator. However, the GPT-3 translation uses a more general term "The absolute soul" which doesn't accurately reflect the Arabic term.

Fluency. : the human translation is more fluent and reads more naturally than the GPT-3 translation. For instance, the phrase "حباه الله بها" (habbahu Allahu biha) in the

Arabic text is translated as "a gift that has been bestowed upon Allah's messenger"

in the human translation. This phrase is more fluent and appropriate in English than the GPT-3 translation, which uses the phrase "which Allah has blessed him with".

Cohesion and coherence. : the human translation is more cohesive and coherent than the GPT-3 translation. The human translator has managed to convey the intended meaning of the original Arabic text more accurately and smoothly than the machine translation. For example, the phrase " جعلت تجوب الأكوان وترى بديعَ الخلق وعظيمَهَ" (ja'alt tajubu al-akwan wa tara bid'i'a al-khalq wa 'azhima) in the Arabic text is translated as "set it to roam the universe, explore, and hover over the earth to witness Allah's grand creation" in the human translation, which is more coherent and conveys the intended meaning better than the GPT-3 translation, which reads "So Allah granted him his soul and made it roam the universes, seeing the wonders of creation and its greatness, and hovering over the earth to see and hear."

Style and register. : the human translation is more appropriate in terms of style and register for the intended audience. The human translator has managed to convey the intended message in a style that is appropriate for the target audience. For example, the use of the phrase "may Allah's blessings be upon him" in the human translation is a common and appropriate expression of respect for the Prophet Muhammad (PBUH) in Islamic culture. The GPT-3 translation, on the other hand, doesn't include this expression.

Word choice. : the human translation is more accurate in terms of word choice than the GPT-3 translation. The human translator has managed to use words and expressions that convey the intended meaning more accurately than the machine translation. For instance, the use of the phrase "devoted his worldly and divine matters genuinely to Allah" in the human translation conveys the intended meaning more accurately than the GPT-3 translation, which uses the phrase "dedicated everything to Allah".

Original Arabic Text:

. الروح المطلقة، مسخرة لدين الله، أطلقها الله تعالى لترى وتصف فتدلّ عليه جلّ وعلا، وأعطاها القدرات لتُريَ الناس قدرة الله لا قدرة العبد المقتدر، فتؤمن بالله القدير وليس بالعبد الفقير.

Human Translation

. An absolute unconfined soul serves Allah's creed; it describes what it witnesses to reveal Allah's ultimate supremacy verses man's vulnerability, preaching humanity to believe in Him, Almighty. Though the spheres of the absolute soul are numerous, their ultimate focus is saving humans and guiding the conduct of potential believers.

GPT-3 Translation

. The absolute soul is dedicated to the religion of Allah, which Allah has granted it to see and describe, and has given it the ability to show people the power of Allah, not the power of a capable servant, so they believe in Allah the Almighty, not in the poor servant.

Direct Assessment:

Accuracy:. both translations convey the overall meaning of the original text. However, there are some differences in the choice of words and phrasing. For example, the human translation uses "unconfined soul" while the GPT-3 translation uses "absolute soul." Additionally, the human translation uses "saving humans" while the GPT-3 translation uses "show people." These differences may affect the nuanced meaning of the text.

Fluency.: the human translation appears to be more natural and engaging, with clear and concise sentences. The GPT-3 translation has some grammatical errors, such as the twisted sentence structure "not the power of a capable servant, so they believe in Allah the Almighty, not in the poor servant.

Cohesion and coherence: the human translation appears to be better organized, with clear transitions between ideas and a logical flow. The GPT-3 translation has some awkward phrasing and does not flow as smoothly.

Style and register.: the human translation uses a more formal and religious tone, while the GPT-3 translation is more straightforward and less poetic.

Word Choice.: the human translation appears to be more precise and thoughtful in the selection of words, while the GPT-3 translation uses more general and commonly used words.

Original Arabic Tex

الحمد لله جامع الناس في الدنيا وفي الآخرة، القائل: {يَا أَيُّهَا النَّاسُ إِنَّا خَلَقْنَاكُمْ مِنْ ذَكَرٍ وَأُنْثَى وَجَعَلْنَاكُمْ شُعُوبًا وَقَبَائِلَ لِتَعَارَفُوا إِنَّ أَكْرَمَكُمْ عِنْدَ اللَّهِ أَتْقَاكُمْ إِنَّ اللَّهَ عَلِيمٌ خَبِيرٌ} [الحجرات: 13]. والصلاة والسلام على سيدنا محمد الذي أوضح معاني هذا التعارف ومنها قوله: " إِنَّ اللَّهَ فَرَضَ عَلَى أَغْنِيَاءِ الْمُسْلِمِينَ فِي أَمْوَالِهِمْ بِقَدْرِ الَّذِي يَسَعُ فُقَرَاءَهُم".

Human Translation

. Praise be to Allah, The Gatherer of people in this world and in the hereafter, The One Saying: “O humanity! Indeed, We created you from a male and a female, and made you into peoples and tribes so that you may ˹get to˺ know one another. Surely the most noble of you in the sight of Allah is the most righteous among you. Allah is truly All-Knowing, All-Aware” [49:13]. And prayers and peace be upon our Master Mohammad who explained Man’s obligations towards this assembly, one of which is practicing charitable giving. He, peace be upon him, said: “Allah has enjoined upon the rich Muslims in their wealth in proportion to that which suffices their poor”.

GPT-3 Translation

. Praise be to Allah, the One who gathers people in this world and in the Hereafter, as He says: "O mankind, indeed We have created you from male and female and made you peoples and tribes that you may know one another. Indeed, the most noble of you in the sight of Allah is the most righteous of you. Indeed, Allah is Knowing and Acquainted." [Quran 49:13] And peace and blessings be upon our Prophet Muhammad, who clarified the meanings of this mutual recognition, including his statement: "Indeed, Allah has made it obligatory upon the rich among the Muslims to give a portion of their wealth to the poor according to what will suffice them."

Direct Assessment:

Accuracy. : The human translation accurately conveys the meaning of the original Arabic text without any major deviation. The GPT-3 translation also conveys the meaning of the text but has a minor deviation in the last sentence where it translates it inaccurately as "a portion of their wealth to the poor according to what will suffice them".

Fluency. : The human translation is fluent and reads naturally in English. The GPT-3 translation is also fluent but has some grammatical errors such as using "gathers" instead of "The Gatherer".

Cohesion and coherence. : The human translation has a coherent flow and maintains the connections between the sentences, with appropriate use of transitional words. The GPT-3 translation has a generally coherent flow, but lacks the appropriate use of transitional words, which may affect the overall coherence.

Style and register. : The human translation uses appropriate style and register, with proper use of formal language. The GPT-3 translation also uses formal language, but may not be as appropriate in style and register as the human translation.

Word choice. : The human translation uses appropriate word choices and maintains the meaning of the original text. The GPT-3 translation has some instances of incorrect word choice such as "including his statement".

Original Arabic Text:

. روح العطاء تحذّر الذين يحجُبون الصدقة عن مستحقيها ويبقونها في أموالهم، أنها جمرةٌ قد تحرق مالهم كلَّه، جاء في الحديث القدسي: " يَا ابْنَ آدَمَ، أَوْدِعْ مِنْ كَنْزِكَ عِنْدِي لَا حَرَقَ، وَلَا غَرقَ، وَلَا سَرَقَ أُوفِيكَهُ أَحْوَجَ مَا تَكُونُ إِلَيْهِ "، وقَالَ رَسُولُ اللهِ، ﷺ ، في ذلك:
" صَنَائِعُ الْمَعْرُوفِ تَقِي مَصَارِعَ السَّوْءِ , وَالْآفَاتِ وَالْهَلَكَاتِ،..".

Human Translation

. The soul of charity warns those who hold charity from its beneficiaries and keep belongings to themselves from losing all their money. Allah says in the Holy Hadith: "O son of Adam, entrust some of your treasure with me that is neither burnt, nor drowned, nor stolen" . On that prophet Mohammad, peace be upon him, also said: “Good deeds protect against evil afflictions, pests and dooms” . He, peace be upon him, further promised believers that: “Indeed charity extinguishes Allah's anger, and it protects against the evil death” . He, peace be upon him, likewise said: “Handling the poor person prevents bad endings” , most of which is Man’s dreadful doom.

GPT-3 Translation

. The spirit of giving warns those who withhold charity from its rightful recipients and keep it in their wealth, that it is a fire that may burn all their wealth. It was mentioned in the Hadith Qudsi: "O son of Adam, deposit your treasure with Me, it will neither be burned nor drowned nor stolen. I will give it back to you in your time of need." And the Messenger of Allah, peace be upon him, said: "Good deeds protect against misfortunes, calamities, and destruction."

Direct Assessment:

Accuracy. : Both the human translation and the GPT-3 translation accurately captures the meaning of the original Arabic text.

Fluency. : The human translation is more fluent and natural than the GPT-3 translation. The GPT-3 translation contains some awkward phrasings and unnatural syntax, which detracts from the readability of the text.

Cohesion and coherence. : The human translation is more cohesive and coherent than the GPT-3 translation. The GPT-3 translation jumps between different topics without proper transition, making the text disjointed and confusing.

Style and register. : The human translation demonstrates a deeper understanding of the style and register of the original Arabic text. The GPT-3 translation lacks the appropriate register and tone for religious texts, which may make it less effective in conveying the intended message.

Word choice. : The human translation uses more precise and appropriate word choices compared to **the GPT-3 translation.** . For example, the human translation uses "beneficiaries" instead of "rightful recipients," which is a more accurate translation of the word "مستحقيها" in the original Arabic text.

Based on the analysis conducted, the human translation appears to outperform the GPT translation in terms of accuracy, fluency, cohesion and coherence, style and register, and word choice. While the GPT translation produced a generally understandable version of the original Arabic text, it often failed to accurately capture the nuances and subtleties of the language and context. The human translation demonstrated a deeper understanding of the language and cultural context, resulting in a more natural and engaging rendition of the text. However, it is worth noting that the GPT translation has the advantage of being produced quickly and without the need for human intervention, making it a useful tool for certain applications where speed is more important than precision. Overall, while machine translation technology has made significant progress in recent years, it still falls short of the level of quality and accuracy that can be achieved by human translators.

Measuring Translation Quality: Quantitative Assessment

. In order to assess the quality of the translations generated by GPT-3 compared to human translations, we utilized quantitative evaluation metrics such as BERT and Rouge scores. These metrics allowed us to objectively compare the translations generated by GPT-3 to those produced by human translators and identify areas where GPT-3 excelled or fell short. In this subsection, we will delve into the details of how we used these metrics to evaluate the translation quality of GPT-3 and provide insights into the strengths and weaknesses of this language model in the context of machine translation.

Comparison of ROUGE and BERT Performance using Bar Graphs

. Looking at the Figure 4, we can see that BERT scores are consistently higher than ROUGE scores for all chapters. This indicates that BERT is a more accurate metric for evaluating the quality of translations produced by GPT-3, as compared to ROUGE. The mean BERT score across all chapters is 0.828, which is significantly higher than the mean ROUGE score of 0.130. This suggests that the translations produced by GPT-3 are of a relatively high quality, especially considering the complexity of the specialized text being translated.

Furthermore, it is interesting to note that the BERT scores for each chapter are relatively consistent, with only a small range of scores observed across chapters. This indicates that GPT-3 is able to maintain a high level of translation quality across different types of specialized text. On the other hand, the ROUGE scores show more variation across chapters, with some chapters having much lower scores than others. This suggests that ROUGE may not be as reliable a metric for evaluating the quality of translations of specialized text.

Overall, our findings suggest that GPT-3 is capable of producing high-quality translations of specialized text, particularly when evaluated using the BERT metric. However, it is important to note that GPT-3 is not a perfect translator, and there may still be room for improvement in terms of accuracy and fluency. Nonetheless, the results of our study provide valuable insights into the potential uses of GPT-3 in the field of translation, particularly for specialized text.

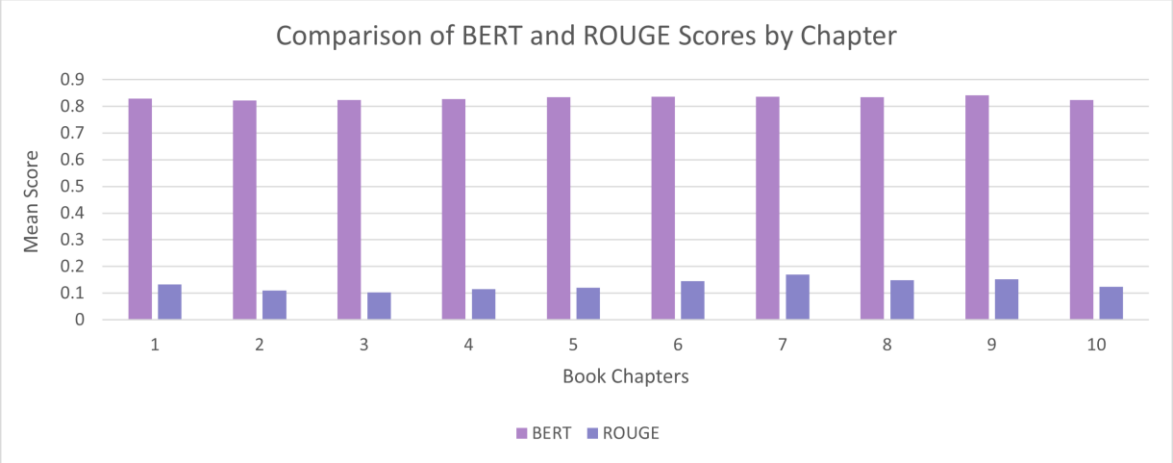


Figure 4. Comparison of BERT and ROUGE scores for specialized text translation

Comparison of ROUGE and BERT Performance using Boxplots

Boxplots, also known as box-and-whisker plots, are a powerful graphical tool used to display the distribution of data. In a boxplot, the data is summarized in a way that shows the median, quartiles, and outliers of the distribution. The box itself represents the interquartile range (IQR), which is the range between the first and third quartiles, and it contains 50% of the data. The median is represented by a line inside the box, and the whiskers represent the range of the data outside the IQR. Outliers, which are data points that are significantly different from the rest of the data, are plotted as individual points [18].

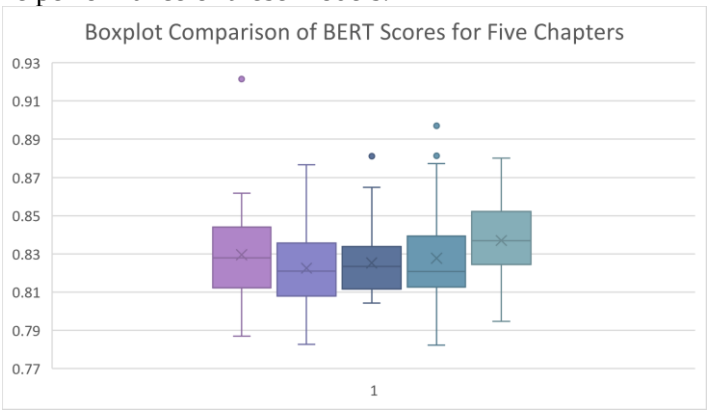
In this study, we employed boxplots as a powerful visualization tool to compare the performance of BERT and ROUGE in five different chapters of a book. To achieve this, we first calculated the BERT and ROUGE scores for sentences in each chapter and then plotted the results as boxplots. Boxplots are a crucial tool for comparing the performance of different evaluation metrics across datasets or experimental conditions, allowing us to gain insights into the central tendency, spread, and variability of the metrics. Through a comparison of the median, interquartile range, and range of the boxplots, we can make informed decisions about which method is better suited for our task. Furthermore, boxplots can highlight potential outliers and skewness in the data, which can significantly impact the validity and reliability of our results. The use of boxplots in our study provided us with an intuitive understanding of the performance of BERT and ROUGE, enabling us to make informed choices and enhance the rigor and credibility of our research.

As seen in Figures 5a and 5b, we can conclude that the performance of BERT in scoring the sentences in the five chapters is more consistent compared to ROUGE. This is indicated by the small variability in the means and standard deviations across the chapters for BERT. Additionally, there are only three outliers present in the BERT boxplot, which indicates that the model's performance is consistent across most of the sentences in the dataset.

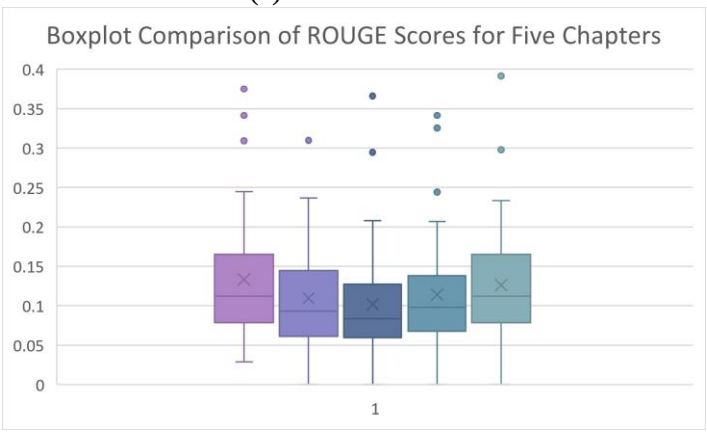
On the other hand, the performance of ROUGE in scoring the sentences in the five chapters is less consistent compared to BERT. This is indicated by the larger variability in the means and standard deviations across the chapters for ROUGE. Furthermore, there are 11 outliers present in the ROUGE boxplot, which indicates that the model's performance is highly variable across the sentences in the dataset.

Overall, the boxplots suggest that BERT is a more consistent and reliable model for scoring sentences compared to ROUGE. However, it's important to note that this conclusion is based solely on the

performance of the models on this particular dataset, and may not necessarily hold true for other datasets or tasks. Further analysis and experimentation would be required to draw more general conclusions about the performance of these models.



(a) BERT Score



(b) ROUGE Score

Figure 5. Comparison of BERT and ROUGE scores for 5 chapters

Paired T-test

. The paired t-test results for BERT and ROUGE scores are presented in Table 1. The table shows the *t*-statistic and *p*-value for each chapter of the text, with a total of 10 chapters analyzed.

Overall, the results show that there is a statistically significant difference between the BERT and ROUGE scores for all chapters analyzed. The *p*-values for each chapter are very small, ranging from 7.96115e-57 to 1.26939e-37, indicating strong evidence against the null hypothesis of no difference between the two scores. The *t*-statistics are all quite large, with values ranging from 41.8546 to 89.3597, further supporting the conclusion of a significant difference. This suggests that BERT performs better than ROUGE in this context, and may be a more reliable and accurate method for evaluating text translation systems.

It is also interesting to note the variability in the results across different chapters. For example, the *t*-statistic is highest for chapter 5, indicating the largest difference between the two scores for that chapter. On the other hand, the *t*-statistic is lowest for chapter 7, indicating a smaller difference between the two scores. This variability may be due to differences in the text content, writing style, or other factors that could affect the performance of the BERT and ROUGE algorithms.

We can conclude that BERT is more robust than ROUGE for Arabic translation. The *t*-statistics for all chapters indicate that the difference in BERT scores between the original and translated text is statistically significant at a very high confidence level, with *p*-values ranging from 7.96115×10^{-57} to 1.26939×10^{-37}.

One potential explanation for this difference in performance is that BERT is able to capture the meaning of the text better than ROUGE, as it considers the context and surrounding words when making predictions. This is particularly important in Arabic, which has a rich morphology and a complex grammar that can lead to multiple valid interpretations of a sentence. BERT's ability to look at the content of the text rather than just the individual words themselves allows it to better understand the underlying meaning of the text and produce more accurate translations.

These results suggest that BERT is a more suitable model for Arabic translation tasks, especially when dealing with complex or ambiguous sentences. One potential explanation for this difference in performance is that BERT is able to capture the meaning of the text better than ROUGE, as it considers the context and surrounding words when making predictions. This is particularly important in Arabic, which has a rich morphology and a complex grammar that can lead to multiple valid interpretations of a sentence. BERT's ability to look at the content of the text rather than just the individual words themselves allows it to better understand the underlying meaning of the text and produce more accurate translations.

Table 1. Paired t-test results for BERT and ROUGE scores

Chapter	Paired t-test results	
	t-statistic	*p*-value
1	63.0107	1.26939e-37
2	80.211	3.46191e-42
3	81.6438	5.16161e-47
4	67.6091	1.56427e-39
5	89.3597	7.96115e-57
6	48.402	8.40789e-38
7	41.8546	2.11931e-33
8	44.0869	5.68029e-48
9	61.8049	6.83792e-49
10	78.9156	3.64275e-48

5. Conclusion and Future Perspectives

The study conducted on the translation capabilities of GPT-3 with respect to specialized religious text has shown promising results. However, it is important to note that further research is required to fully understand and address the limitations of the model in this area. While GPT-3 has demonstrated a reasonable level of accuracy and fluency in translating religious texts, it is essential to improve its performance and accuracy. Future research could focus on improving the training data used to train machine translation models to better handle domain-specific vocabulary and terminology, as well as developing new evaluation metrics that can accurately measure the quality

of the translations. Additionally, further efforts could be directed towards enhancing the model's ability to handle complex sentence structures and discourse phenomena commonly present in religious texts. By addressing these issues, GPT-3 could potentially become a valuable tool for translating specialized religious text.

In addition to the potential benefits for religious institutions, the findings of this study could have far-reaching implications for various industries and fields that rely heavily on specialized text translation, such as legal, medical, and technical domains. The ability of machine translation systems to accurately and efficiently handle specialized text can significantly reduce the time and costs associated with manual translation, improving overall efficiency and productivity. For example, in the legal field, accurate and efficient translation of legal documents can play a critical role in ensuring fairness and justice for non-native speakers. Similarly, in the medical field, machine translation can help facilitate communication between doctors and patients from different linguistic backgrounds, potentially improving patient outcomes. The development of machine translation systems that can effectively handle specialized text has the potential to revolutionize various industries and improve global communication and collaboration.

In conclusion, this study has emphasized the importance of ongoing research and development in the field of machine translation of specialized text. The promising potential of GPT-3 and other NLP models in this area offers exciting opportunities for the creation of more advanced and accurate machine translation systems that can meet the increasing demand for specialized text translation in various industries and fields. The future of machine translation appears to be bright, and with further research and development, it could become an indispensable tool in many industries, enhancing efficiency, productivity, and accuracy. As such, this study could inspire researchers and practitioners to explore the possibilities of machine translation further and find innovative ways to enhance its capabilities. Ultimately, the development of advanced machine translation systems can contribute to a more connected world, bridging language barriers and facilitating communication and understanding across different cultures and regions.

References

[1] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, ``How good are gpt models at machine translation? a comprehensive evaluation,'' 2023.

[2] Z. Tan, S. Wang, Z. Yang, G. Chen, X. Huang, M. Sun, and Y. Liu, ``Neural machine translation: A review of methods, resources, and tools,'' AI Open, vol. 1, pp. 5--21, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/ pii/S2666651020300024

[3] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, ``Bertscore: Evaluating text generation with bert,'' 2020.

[4] C.-Y. Lin, ``Rouge: A package for automatic evaluation of summaries,'' in Text summarization branches out, 2004, pp. 74--81.

[5] L. Zhou, W. Hu, J. Zhang, and C. Zong, ``Neural system combination for machine translation,'' arXiv preprint arXiv:1704.06393, 2017.

[6] P. Koehn and R. Knowles, ``Six challenges for neural machine translation,'' arXiv preprint arXiv:1706.03872, 2017.

[7] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, ``Gpt understands, too,'' arXiv preprint arXiv:2103.10385, 2021.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan,

[9] P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, ``Language models are few-shot learners,'' 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ``Attention is all you need," 2017.
[11] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al., ``Improving language understanding by generative pre-training," 2018.
[12] W. Jiao, W. Wang, J. tse Huang, X. Wang, and Z. Tu, ``Is chatgpt a good translator? yes with gpt-4 as the engine," 2023.
[13] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, and Z. Tu, ``Document-level machine translation with large language models," 2023.
[14] S. Castilho, C. Mallon, R. Meister, and S. Yue, ``Do online machine translation systems care for context? what about a gpt model?" 2023.
[15] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, ``Multilingual machine translation with large language models: Empirical results and analysis," 2023.
[16] E. Chatzikoumi, ``How to evaluate machine translation: A review of automated and human metrics," Natural Language Engineering, vol. 26, no. 2, pp. 137--161, 2020.
[17] M. Farshoukh, Soul Breezes. Beirut: Iijazforum, 2018.
[18] T. K. Kim, ``T test as a parametric statistic," Korean journal of anesthesiology, vol. 68, no. 6, pp. 540--546, 2015.
[19] D. Williamson, R. Parker, and J. Kendrick, ``The box plot: A simple visual method to interpret data," Annals of internal medicine, vol. 110, pp. 916--21, 07 1989.