information

Article

Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems

Jungha Son and Boyoung Kim *

Seoul Business School, aSSIST University, Seoul 03767, Republic of Korea; jhson@stud.assist.ac.kr
* Correspondence: bykim2@assist.ac.kr; Tel.: +82-10-4046-2428

Abstract: The rapid global expansion of ChatGPT, which plays a crucial role in interactive knowledge sharing and translation, underscores the importance of comparative performance assessments in artificial intelligence (AI) technology. This study concentrated on this crucial issue by exploring and contrasting the translation performances of large language models (LLMs) and neural machine translation (NMT) systems. For this aim, the APIs of Google Translate, Microsoft Translator, and OpenAI's ChatGPT were utilized, leveraging parallel corpora from the Workshop on Machine Translation (WMT) 2018 and 2020 benchmarks. By applying recognized evaluation metrics such as BLEU, chrF, and TER, a comprehensive performance analysis across a variety of language pairs, translation directions, and reference token sizes was conducted. The findings reveal that while Google Translate and Microsoft Translator generally surpass ChatGPT in terms of their BLEU, chrF, and TER scores, ChatGPT exhibits superior performance in specific language pairs. Translations from non-English to English consistently yielded better results across all three systems compared with translations from English to non-English. Significantly, an improvement in translation system performance was observed as the token size increased, hinting at the potential benefits of training models on larger token sizes.

Keywords: large language model; neural machine translation; ChatGPT; Google Translate; Microsoft Translator

Citation: Son, J.; Kim, B. Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. *Information* **2023**, *14*, 574. https://doi.org/10.3390/info14100574

Academic Editor: Katsuhide Fujita

Received: 31 July 2023
Revised: 3 October 2023
Accepted: 18 October 2023
Published: 19 October 2023

1. Introduction

The emergence of LLMs, such as ChatGPT, has recently become the focus point in the AI industry, prompting the language services sector to explore the implications of these multifunctional LLMs capable of assisting in various tasks, including text generation, summarization, translation, and code generation [1]. According to a report published by Acumen Research and Consulting, the global machine translation market size was USD 812.6 million in 2021 and is expected to reach USD 4069.5 million by 2030, growing at a CAGR of 19.9% from 2022 to 2030 [2].

The accelerated progress in AI and natural language processing (NLP) has not only fostered the development of highly sophisticated and adaptable language models [3–5] but has also exerted a considerable influence on numerous domains, particularly machine translation, where traditional NMT techniques have achieved remarkable advancements in recent years [6].

Several key approaches have shaped the evolution of machine translation. Initially, rule-based machine translation (RBMT) utilized grammar rules and lexicon dictionaries but faced limitations due to computational power and data scarcity. Statistical machine translation (SMT) later emerged, leveraging large parallel corpora to improve translation performance [7]. NMT then incorporated deep learning and artificial neural networks, with the transformer model becoming particularly notable [8]. Hybrid machine translation methods have combined rule-based, statistical, and neural techniques to enhance translation quality [9].

Pre-trained language models such as GPTs (generative pre-trained transformers) and BERT (bidirectional encoder representations from transformers) have significantly impacted NLP and demonstrated excellent performance in diverse NLP tasks, including translation, due to their self-supervised learning, transfer learning, and adaptability to different contexts [10,11]. This progress has led to research focusing on model scalability, efficiency, domain adaptability, and human-like understanding and translation capabilities. However, challenges remain, such as data bias, model interpretability, and energy efficiency [12].

As the field of machine translation continues to progress, it is anticipated that systems will become more accurate, robust, and adaptable to a wider range of languages and domains, making them even more valuable tools in our interconnected world. This research aimed to compare the translation performance of LLMs with that of traditional NMT models from the perspective of real-world users. By conducting a comprehensive evaluation, this study provides valuable insights into the practicality and effectiveness of LLMs in translation tasks, highlighting their strengths and weaknesses compared with NMT approaches.

Considering fast-evolving generative AI services and language-based AI technology usability, this study presents a range of techniques for language translation requiring accuracy, as well as professional experimental results for data usage. This has the significance of research to look at better language translation problems in different languages from the different technologies of LLM. Furthermore, given the complexity and rapid changes in the OpenAI ecosystem, it will be necessary to discuss ecosystem expansion and development via learning about LLMs and API integration. LLM will also contribute to creating a more advanced discourse by creating a foundation for global cooperation among experts beyond the open sourcing of technology and data in the future, as it requires efforts to expand the ecosystem based on cooperation with global networks rather than specific experts or organizations.

In light of the aforementioned, the pivotal contributions of this paper unfold in three dimensions. First, a meticulous exploration of the translation performance of prominent LLMs and NMT systems is undertaken, providing a comprehensive analysis across various language pairs and translation directions. Second, an in-depth investigation into the influence of token size on translation quality is conducted, elucidating the potential advantages of training models with larger token sizes. Lastly, the translation capabilities of ChatGPT in specific language pairs are scrutinized, uncovering its competitive performance in certain contexts.

The remainder of this paper is organized as follows: Section 2 provides a theoretical background on machine translation, emphasizing the metrics used for evaluating translation performance. Section 3 describes the experimental setup, detailing the selection of test subjects and data. Section 4 presents and analyzes the experimental results. Section 5 offers conclusions, summarizing the main findings and suggesting potential avenues for future research.

2. Theoretical Background

2.1. Machine Translation Approaches and Evolution

Machine translation refers to the process of automatically translating text from one natural language to another using computer algorithms [13]. Numerous approaches have been developed to tackle the inherent challenges in translating complex and diverse human languages [14]. Depending on the subject, translation can be classified into human translation, human-assisted translation (HAMT), and computer-assisted translation (CAT) according to the degree of human intervention [7,15,16]. With the development of AI technology, interest in artificial neural network machine translation has increased. However, although the quality of machine translation has significantly improved, the quality of human translation remains superior [17].

Three machine translation technologies have been developed so far. First, RMBT relies on manually created linguistic rules and dictionaries to map the source language's grammar,

syntax, and semantics onto the target language [7,18]. RBMT systems can be further divided into transfer-based, interlingua-based, and direct translation models. However, due to the inherent complexity of languages and the labor-intensive process of rule creation, RBMT has limitations in scalability and adaptability.

Second, SMT builds probabilistic models based on large-scale parallel corpora to perform translations [19]. SMT models learn the most probable translation of a source sentence by finding the best alignment of words or phrases between the source and target languages. Although SMT has shown significant improvements compared with RBMT and example-based machine translation (EBMT), it still faces issues in handling long-range dependencies and capturing context.

Third, EBMT generates translations by searching for matching patterns in a database of previously translated sentences or phrases [20]. EBMT systems then apply various algorithms to combine and adapt the matched examples to produce the target translation. While EBMT can produce more fluent translations than RBMT, it is still limited by the availability of suitable examples in the database.

However, with the advent of deep learning techniques, research on NMT has gained prominence. NMT models, such as the encoder–decoder architecture and the attention mechanism, have demonstrated significant improvements in translation quality by capturing long-range dependencies and handling context better than previous approaches. Moreover, the transformer model has significantly enhanced NMT performance and strengthened parallel processing and interlayer communication [8,21].

Hybrid systems that integrate NMT, SMT, or RBMT methods have also been investigated. These hybrid systems leverage the strengths of both approaches to address some of the limitations of individual methods, resulting in higher-quality translations. Each approach has its strengths and weaknesses, and ongoing research aims to address these limitations to improve translation quality and applicability across diverse languages and domains. Traditional NMT models primarily translate sentences in isolation, neglecting the broader document context. However, recent research has emphasized document-level machine translation that considers the overall document context. This burgeoning field involves novel modeling, architectures, training, and decoding strategies and introduces evaluation metrics and discourse-focused test sets to measure improvements in document-level machine translation [22,23].

The development and success of LLMs in recent years have led to a surge in interest in their application to machine translation tasks. LLMs have demonstrated their ability to generalize across various NLP tasks, thanks to self-supervised learning and transfer learning, which have resulted in improved translation performance. In addition, BERT has been incorporated into NMT models as a pre-training mechanism, leading to improved translation quality in various settings [24]. Moreover, several BERT-based models, such as mBERT (multilingual BERT) and XLM-R (Cross-lingual Language Model—RoBERTa), have been developed to handle multilingual and cross-lingual tasks [25,26].

OpenAI's GPT-2, GPT-3, and GPT-4 are examples of LLMs that have significantly impacted the NLP field, including machine translation [27–29]. GPT-3, in particular, has demonstrated remarkable performance in zero-shot machine translation tasks, where the model translates text without being specifically fine-tuned for translation [28]. GPT-4 is the latest iteration of GPT, and is said to be even more powerful and capable than its predecessors. It has been trained on a massive dataset of text and code, and can generate text, translate languages, write different kinds of creative content, and answer questions in an informative way [30].

The utilization of LLMs in machine translation has received significant attention in recent years, and research has aimed to compare the performance of LLMs and traditional NMT systems. LLMs such as GPT and BERT have exhibited notable enhancements in translation performance compared with traditional NMT systems [27].

2.2. Metrics for Evaluating Machine Translation Performance

The concept of machine translation has existed since the 1950s and was initially predicated on a rule-based approach using linguistic rules and bilingual dictionaries for translation [13]. Nevertheless, evaluating these systems was a challenging task due to the absence of standard measures [31]. The introduction of the BLEU (bilingual evaluation understudy) metric marked a significant advancement in addressing this challenge and a landmark breakthrough in the evaluation of machine translation systems [32]. Various metrics have been proposed to assess the quality of machine translation output, including BLEU, chrF (character n-gram F-score), and TER (translation edit rate). These metrics are widely used in research and industry due to their effectiveness and ease of use.

Papineni et al. [32] first proposed the BLEU metric as a corpus-based metric for evaluating the quality of machine translation output against human reference translations. The BLEU metric is computed as follows:

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} \omega_n log P_n\right) \tag{1}$$

$$\begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases} \tag{2}$$

where P_n represents the precision of n-grams, ω_n is the weight of the precision (often set to $\frac{1}{N}$ for uniformity), BP is the brevity penalty, c is the length of the candidate translation, and r is the length of the closest reference translation.

BLEU measures the n-gram overlap between the machine-generated translation and reference translations, where a higher score indicates greater similarity to the human reference. BLEU has become one of the most widely used metrics for evaluating machine translation performance and has been shown to correlate well with human judgments of translation quality. Although the BLEU metric introduced a simple and effective way to evaluate machine translation output using reference translations [32], it is not without its limitations, as identified in subsequent research. Among these limitations is the lack of consideration for morphological and syntactic variations, which are significant aspects of certain languages, as well as the absence of recall measurement [33].

In response to these challenges, Popović [34] introduced the chrF metric. This metric computes the character n-gram precision and recall, typically defined as

$$chrF = \beta^2 \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{3}$$

where β weighs the emphasis of recall over precision, and is defined as

$$Precision = \frac{Number\ of\ correct\ character\ n-grams}{Total\ number\ of\ character\ n-grams\ in\ the\ generated\ text} \tag{4}$$

$$Recall = \frac{Number\ of\ correct\ character\ n-grams}{Total\ number\ of\ character\ n-grams\ in\ the\ reference\ text} \tag{5}$$

This makes chrF adept at considering the morphological richness of languages, thereby rendering it an effective evaluation tool, especially for languages with complex morphologies and syntax.

TER was first proposed by Snover et al. [35] as a measure of the edit distance between the machine-generated translation and the human reference translation. Mathematically, the TER metric is defined as

$$TER = \frac{Number\ of\ Edits}{Average\ Length\ of\ Reference\ Translation} \tag{6}$$

It quantifies the number of edits required to transform the machine translation into the reference translation, providing insights into the fluency and adequacy of the translation. The TER metric has been widely adopted and provides a complementary perspective on the quality of machine translation systems, considering the specific edits made during the translation process. It offers valuable information, particularly in cases where the translations may differ significantly from the reference translations, helping to assess the overall accuracy and fidelity of the machine-generated output.

While metrics such as BLEU, chrF, and TER significantly contribute to the aim of quantifying machine translation quality, crucially, they do not fully capture all the aspects of translation quality [36]. They tend to primarily focus on surface-level linguistic matches between the machine output and human reference translations, while elements such as fluency, idiomaticity, and cultural appropriateness, among others, are not directly measured [37]. Consequently, these metrics should be used in conjunction with other evaluation approaches, including human judgment, to provide a more comprehensive evaluation of machine translation systems [38].

3. Method

3.1. Selection of Test Subjects

3.1.1. Selection of Test Systems

To evaluate a diverse range of language pairs and translation directions, encompassing various linguistic families and language complexities, Google Translate and Microsoft Translator were selected as representative NMT systems to assess and compare their performance with that of OpenAI's ChatGPT, a prominent large language model (LLM).

As shown in Table 1, Google Translate and Microsoft Translator are state-of-the-art NMT systems adopted by users worldwide [39,40]. These systems leverage advanced deep learning techniques, such as the encoder–decoder architecture and attention mechanisms, which enable them to improve translation quality by effectively capturing long-range dependencies and more accurately handling context compared with earlier methods [41]. The encoder–decoder architecture is the foundation of these NMT systems, with the encoder processing the input text in the source language and the decoder generating the output text in the target language [42].

Table 1. Selected evaluation systems.

Section	System Characteristics
Google Translate	Google Translate is a widely used NMT system that leverages Google's extensive research in the field of machine translation. It supports a vast number of languages and has been continually improved over the years, resulting in high-quality translations across various language pairs.
Microsoft Translator	Developed by Microsoft Azure Cognitive Services, Microsoft Translator is another leading NMT system that offers translation capabilities for numerous languages. It employs advanced neural network techniques and benefits from Microsoft's extensive research on machine translation, providing accurate translations for a diverse set of language pairs.
ChatGPT	ChatGPT is a family of advanced LLMs developed by OpenAI to excel in tasks such as conversation, question answering, and content generation. ChatGPT has also shown remarkable efficacy in translation applications despite not being initially designed for this purpose.

Attention mechanisms play a crucial role in the translation process by allowing the model to weigh the importance of different parts of the input text when generating the output. This approach addresses the limitations of previous models that struggled to maintain context and coherence in longer sentences [8]. ChatGPT is built upon the advanced GPT architecture, which has shown exceptional performance across a wide range of NLP tasks, including translation [43]. The accuracy of this model in providing information or answers has been considerably enhanced via key innovations such as large-scale pre-training, instruction fine-tuning, and reinforcement learning from human feedback (RLHF) [44].

The model's capacity to generalize and adapt to diverse contexts stems from its use of self-supervised learning and transfer learning techniques. These features have garnered significant attention, highlighting ChatGPT's potential as a powerful translation tool [45]. The underlying GPT architecture utilizes a transformer-based design, enabling it to capture long-range dependencies and contextual information more effectively than previous models. Furthermore, ChatGPT benefits from the vast amount of training data and extensive pre-training, which allow it to generate highly accurate and fluent translations, even for complex sentences and specialized domains [29].

Hence, this study utilized the APIs of Google Translate, Microsoft Translator, and OpenAI's ChatGPT to conduct translation tests on large-scale data across various parallel corpora. The use of these APIs enabled the efficient and accurate evaluation of the translation performance of each system under consideration.

3.1.2. Selection of Metrics and Libraries

Various evaluation metrics were utilized to thoroughly assess the translation performance of ChatGPT, Google Translate, and Microsoft Translator. These metrics comprised BLEU, chrF, and TER scores. The BLEU, chrF, and TER scores for translations generated by ChatGPT, Google Translate, and Microsoft Translator were calculated using the 'SacreBLEU library'. To maintain consistency across all experiments, we used the default parameters provided by the library. The greatest advantage of SacreBLEU is its ability to mitigate the impact of preprocessing variations on score calculations, which can vary with minor changes, such as via tokenizers, the segmentation of compound nouns, the handling of unknown tokens for rare words, or casing. By incorporating text normalization stages in the architecture, this automated metric can provide more reliable evaluation scores [46].

To enable a more straightforward comparison of translation performance across all metrics, the scores were standardized on a consistent 0–100 scale. This facilitates the interpretation and evaluation of the results. For the TER metric, which is originally expressed as an edit rate, the scores were reversed to align with the other evaluation metrics, whereby higher values now indicate better translation quality. This standardization ensures that all metrics provide a unified representation of translation quality and facilitates a comprehensive assessment of the performance of different translation results.

3.2. Experimental Design

3.2.1. Experimental Process

Ensuring consistency in the parameters employed across all three systems was essential for achieving accurate and reliable outcomes. Each API was employed with similar parameter settings throughout the experimental process, and the best practices and recommendations provided by their respective documentation were strictly adhered to. The translation systems operated under equivalent conditions, effectively minimizing potential discrepancies that could have arisen from differing configurations. After the experiments, maintaining consistent parameter settings across all APIs, a framework was established for evaluating and comparing the translation capabilities of ChatGPT, Google Translate, and Microsoft Translator, ensuring a reliable assessment of their performance (see Figure 1).

3.2.2. Evaluation Dataset

To evaluate the translation performance of ChatGPT, Google Translate, and Microsoft Translator, parallel corpora from the Workshop on Machine Translation (WMT) 2018 and 2020 benchmarks were utilized [47,48]. The Workshop on Machine Translation (WMT), initiated in 2006, holds substantial influence in the machine translation field by facilitating innovation and progress via annual evaluations of machine translation systems [49]. This study utilized a parallel corpus consisting of documents in two languages aligned at the sentence level provided by WMT [50], which is considered a high-quality machine translation system. Datasets include various areas of content, such as news articles and reports.

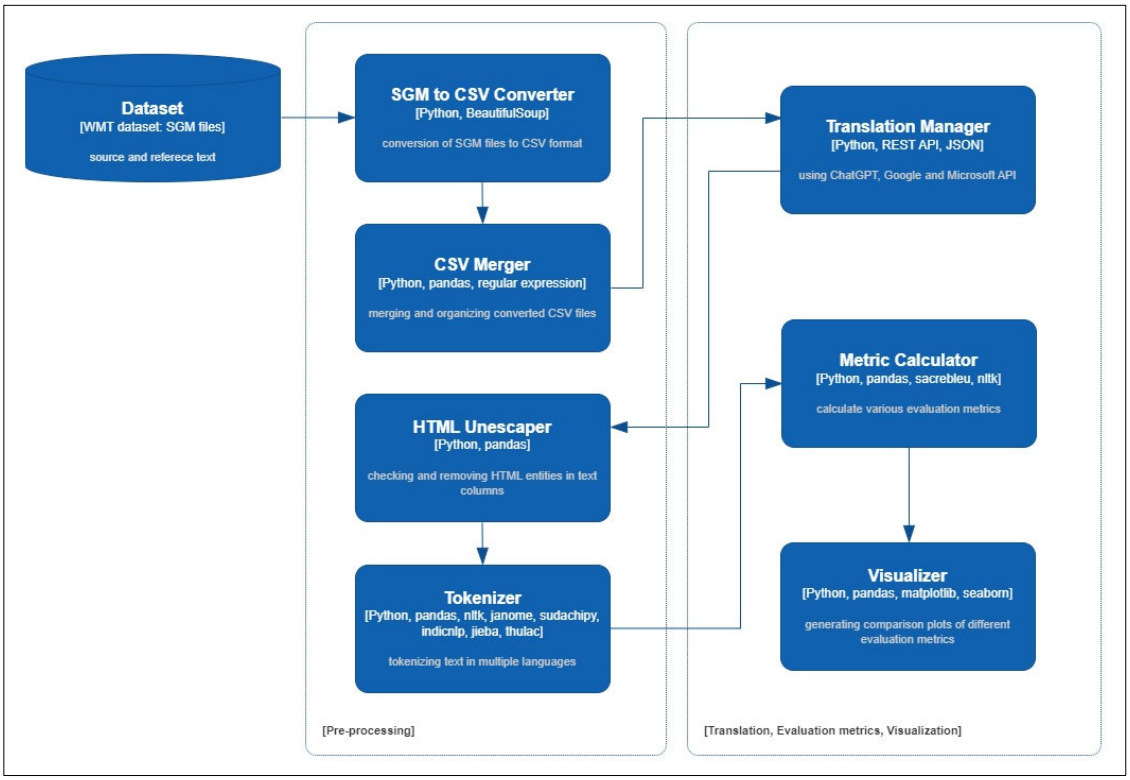

Figure 1. Research design and process.

ChatGPT, Google Translate, and Microsoft Translator then translated the texts between the chosen parallel corpora using these samples, and their performance was measured. In the performance evaluation of the ChatGPT models, four pre-trained models released by OpenAI were compared: GPT-3 (text-curie-001), GPT-3.5 (text-davinci-003), GPT-3.5 (gpt-3.5-turbo), and GPT-4 (limited beta).

We downloaded the language-specific parallel corpora from WMT18 and WMT20 and selected several target languages for our evaluation. The chosen languages represented a diverse range of linguistic families and complexities to ensure a comprehensive assessment of the models' translation capabilities. For this evaluation, 3600 samples were created, with 200 test items selected from each of the 18 language pairs. The translation performance of the GPT models was evaluated on multiple language pairs, including English–Estonian (en-et), Estonian–English (et-en), English–German (en-de), German–English (de-en), English–Finnish (en-fi), Finnish–English (fi-en), English–Russian (en-ru), and Russian–English (ru-en). Each GPT model was utilized to perform translations for these language pairs, and the results were compared with reference translations.

The test dataset was curated to maintain a balance of various text types and domains, ensuring that the models were evaluated on a wide range of translation tasks. The aim of utilizing the WMT18 and WMT20 parallel corpora and making the test dataset was to provide a fair basis for comparing the translation performance of the three models under consideration. In this evaluation, the average scores of the selected parallel corpora were compared for each metric, and the performance was analyzed based on translation direction and sentence length.

Furthermore, the translation performance for each language pair was scored individually, enabling a comparison of translation performance across different language pairs. This test also provides an evaluation of LLMs' ability on discourse based on taking document-level machine translation as a testbed. Because LLMs can produce coherent and relevant answers for various natural language processing (NLP) tasks.

3.2.3. Data Preprocessing and Translation Using API

The parallel corpora were procured from WMT18 and WMT20 and accessed directly from the respective online repositories [47,48]. The downloaded data were in SGML (Standard Generalized Markup Language) format. SGML is a text format used to encode structured text documents, such as those employed in NLP tasks like machine translation. SGML files contain markup tags that indicate the structure of the text, including the beginnings and ends of sentences, paragraphs, and other elements. This format is commonly used for NLP datasets to ensure the consistency and standardization of the data. The SGML files provided by WMT were converted into a CSV format, and the source text was used to call the API of each service to perform translations. The translated results were preprocessed and saved in a CSV format for performance comparison. A large volume of datasets was translated simultaneously by employing Python for automation.

While calling the Google Translation API, it was observed that the text was being escaped using HTML character entities. Consequently, special characters such as quotation marks and apostrophes appeared as HTML character entities in the stored results. To resolve this issue, preprocessing was performed on the translated text using 'Python's html.unescape' method. This method decoded the text by replacing HTML character entities with the corresponding special characters, effectively restoring the original text. The preprocessed text was then saved for subsequent evaluation and analysis (see Figure 2).

Before Pre-processing:

```
&quot;Ocean Change&quot;: Arved Fuchs returns from Greenland expedition
&quot;The Polaroid Diaries&quot; by Linda McCartney
Queen&#39;s granddaughter Princess Beatrice is engaged
&quot;We will continue to buy oil and gas from Iran&quot;
The federal government has now agreed to demand a &quot;declaration of
trust&quot; from suppliers.
EU invests 9.5 million euros in Austria&#39;s railway infrastructure
&quot;The Tears Flowed&quot;
```

After Pre-processing:

```
"Ocean Change": Arved Fuchs returns from Greenland expedition
"The Polaroid Diaries" by Linda McCartney
Queen's granddaughter Princess Beatrice is engaged
"We will continue to buy oil and gas from Iran"
The federal government has now agreed to demand a "declaration of
trust" from suppliers.
EU invests 9.5 million euros in Austria's railway infrastructure
"The Tears Flowed"
```

Figure 2. Examples of preprocessing Google Translation results by removing HTML entities.

This research explored the impact of the reference token size, categorized into quartiles as the average numbers of tokens in a reference text, on the performance of different translation models: GPT-3.5, GPT-4, Google Translate, and Microsoft Translator. By classifying reference texts based on their average token size quartiles, this study shows how token size influences the translation performance of each model. However, in the process of comparing language analysis, unlike many other languages, the Japanese and Chinese languages do not have explicit word boundaries such as spaces, which makes the identification of individual words or tokens crucial for accurate translation evaluation. To address this issue, tokenization was applied to both source and target texts in the Japanese and

Chinese languages before calculating translation quality metrics. Specialized tokenizers were used for each language to handle their intricacies, enabling a more accurate translation quality evaluation. The tokenization process for Japanese and Chinese languages involved segmenting the text into individual tokens or words by analyzing the linguistic structure of the input text, considering elements such as part-of-speech and morphological features. This approach ensured that the translation quality assessment accurately reflected the performance of the respective translation systems for these languages.

4. Results

4.1. Translation Performance Results by chatGPT Models

The performance comparison between the different ChatGPT models on the evaluation dataset is presented in Table 2. The results show a clear progression in the performance of these models, from GPT-3 to the more powerful GPT-3.5 generation models and, ultimately, the GPT-4 model. GPT-4's broader general knowledge and advanced reasoning capabilities contribute to its superior performance in translation tasks, as demonstrated by obtaining the highest BLEU (30.16), chrF (60.84), and TER (42.20) scores. This indicates that as OpenAI continues to develop and refine its models, there is potential for further advancements in translation capabilities. The results obtained from the evaluation led to the selection of GPT-3.5-turbo and GPT-4 as the primary ChatGPT models for comparisons with Google Translate and Microsoft Translator.

Table 2. Performance comparison between ChatGPT models on the evaluation dataset.

Metric	GPT-3 (Text-Curie-001)	GPT-3.5 (Text-Davinci-003)	GPT-3.5 (gpt-3.5-Turbo)	GPT-4 (Limited Beta)
BLEU	13.51	25.33	28.35	30.16
chrF	39.32	56.69	56.67	60.84
TER [1]	19.37	37.14	40.60	42.20

[1] All TER scores throughout this paper are reversed for consistency; higher values represent better quality.

4.2. Translation Performance Results of the Systems

To determine the statistical significance of the observed differences in the translation performance of the systems, the Kruskal–Wallis H test was utilized. This non-parametric method of fitting for the comparison of more than two independent groups was conducted on the BLEU, CHRF, and TER scores. The results of this test are presented in Table 3. *H* test results were significant as BLEU 149.07, chrF 96.34, and TER 225.45. In basic statistics, *df* was 3 for all metrics, and the *p*-Value for all metrics were significant as less than 0.001, implying that the differences in translation performance between the systems are statistically significant.

Table 3. Results of the Kruskal–Wallis H test for translation performance metrics.

Metric	*H* *	*df* **	*p*-Value ***
BLEU	149.07	3	<0.001
chrF	96.34	3	<0.001
TER	225.45	3	<0.001

* H denotes the Kruskal-Wallis H statistic ** df denotes the degrees of freedom *** *p*-Value indicates the significance level.

The translation performance of Google Translate, Microsoft Translator, and ChatGPT was analyzed across 18 language pairs, translating 3600 pieces of data. The performances of the systems on the evaluation dataset are reflected in their respective scores measured with the applied metrics, as presented in Table 4.

Table 4. Performance comparison between the different systems.

Metric	ChatGPT (GPT-3.5)	ChatGPT (GPT-4)	Google	Microsoft
BLEU	24.79	26.86	29.47	29.05
chrF	53.62	55.54	57.38	57.03
TER	36.76	39.10	43.96	43.32

Google Translate and Microsoft Translator outperformed ChatGPT in all the evaluation metrics. Both systems demonstrated higher BLEU, chrF, and TER scores than ChatGPT, indicating better translation qualities. Google Translate demonstrated slightly better performance than Microsoft Translator, but the difference was marginal. Although ChatGPT showed lower performance than the other two systems, it is important to note that ChatGPT is primarily designed as a conversational AI model rather than a translation system. The results suggest that ChatGPT's performance is not on par with that of specialized translation systems such as Google Translate and Microsoft Translator. These findings serve as a valuable reference for users who decide to use translation systems for specific purposes, such that those who prioritize translation quality may opt for Google Translate or Microsoft Translator.

4.3. Translation Performance Comparison for English and Non-English

The translation performance between the English and non-English languages was evaluated in both directions. The results show that translations from non-English to English surpassed those from English to non-English in terms of BLEU, chrF, and TER scores (see Table 5, Figure 3).

Table 5. The comparison results for English and non-English.

Section		ChatGPT (GPT-3.5)	ChatGPT (GPT-4)	Google	Microsoft
BLEU	Non-English–English	26.77	28.71	31.07	29.64
	English–non-English	22.05	24.16	27.54	27.49
chrF	Non-English–English	56.10	57.86	59.35	57.90
	English–non-English	49.52	51.61	54.56	54.81
TER	Non-English–English	38.38	40.62	45.11	43.61
	English–non-English	34.56	36.85	42.61	42.63

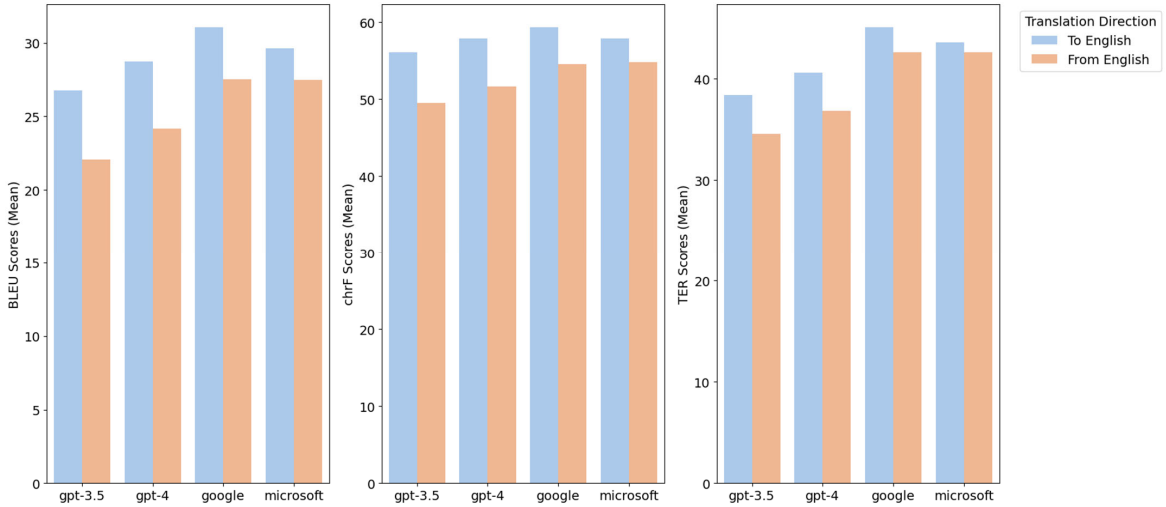

Figure 3. The comparison of scores by translation direction across different metrics.

The average BLEU scores for English–non-English and non-English–English translations show that Google Translate outperforms the other models. In translations from non-English to English, Google Translate achieved the highest average BLEU score of 31.07, surpassing GPT-3.5 with 26.77, GPT-4 with 28.71, and Microsoft Translator with 29.64. Similarly, in translations from English to non-English, Google Translate led with a score of 27.54, followed by Microsoft Translator with 27.49, GPT-4 with 24.16, and GPT-3.5 with 22.05.

Google Translate also excelled in the average chrF scores for both translation directions. In non-English–English translations, Google Translate achieved an average chrF score of 59.35, outperforming GPT-3.5 with 56.10, GPT-4 with 57.86, and Microsoft Translator with 57.90. In English–non-English translations, Microsoft Translator achieved the highest score of 54.81, followed by Google Translate with 54.56, GPT-4 with 51.61, and GPT-3.5 with 49.52. These scores demonstrate Google Translate's superior character-level precision and recall.

Regarding TER scores, Microsoft Translator and Google Translate outperformed ChatGPT in both translation directions. In non-English–English translations, Google Translate achieved the highest average TER score of 45.11, followed by Microsoft Translator with 43.61, GPT-4 with 40.62, and GPT-3.5 with 38.38. Similarly, in English–non-English translations, Microsoft Translator achieved the highest average TER score of 42.63, followed by Google Translate with 42.61, GPT-4 with 36.85, and GPT-3.5 with 34.56. These scores provide valuable insights into the dissimilarity between the translations and the reference translations.

4.4. Translation Performance Comparison by Language Pairs

The results of the evaluation and comparison of GPT-3.5, GPT-4, Google Translate, and Microsoft Translator in terms of their BLEU, chrF, and TER scores for selected language pairs are outlined herein. As seen in Table 6 and Figure 4, the mean BLEU scores of the translation systems reveal a pattern, with performance fluctuating across different language pairs. For instance, in the case of German–English (de-en) translation, GPT-4 exceeded the other systems with a score of 41.23, followed by Microsoft Translator with 40.98, Google Translate with 40.39, and then GPT-3.5 with 37.65.

Table 6. The comparison results by language pairs.

Target	ChatGPT (GPT-3.5)			ChatGPT (GPT-4)			Google			Microsoft		
	BLEU	**chrF**	**TER**	**BLEU**	**chrF**	**TER**	**BLEU**	**chrF**	**TER**	**BLEU**	**chrF**	**TER**
de-en	37.65	66.57	52.93	41.23	68.43	56.39	40.39	67.06	56.07	40.98	67.89	56.16
ru-en	32.91	63.7	46.97	34.67	63.63	48.41	36.41	65.76	51.85	36.5	64.76	51.39
et-en	31.52	63.46	44.76	32.18	64.39	45.52	36.76	65.50	51.40	31.87	66.64	46.28
zh-en	31.48	61.35	43.43	32.13	62.17	43.84	35.13	62.08	47.53	34.97	62.65	48.02
en-de	31.21	61.06	45.70	33.26	62.07	47.46	36.11	64.19	50.51	37.24	60.52	51.27
de-fr	30.52	60.17	45.02	32.51	61.91	47.27	32.23	60.46	47.39	35.41	62.44	48.13
cs-en	29.00	59.98	44.97	30.65	61.96	46.49	29.08	61.12	45.78	28.45	62.30	46.22
en-zh	26.93	59.41	46.23	27.99	60.69	46.08	32.28	58.77	51.39	30.35	58.17	50.36
en-ru	25.84	58.16	38.07	26.76	58.54	37.95	33.40	63.22	47.11	31.55	62.65	46.25
fi-en	25.40	57.11	35.23	25.73	59.51	35.05	25.99	60.03	35.95	27.29	59.80	37.09
fr-de	25.07	55.57	33.20	28.03	58.19	36.88	29.33	60.81	42.12	30.40	61.63	41.71
en-cs	24.92	55.51	38.7	28.63	55.54	42.40	32.01	54.95	47.77	33.53	55.34	48.48
en-et	23.47	54.16	34.46	26.38	56.05	37.73	27.80	56.26	42.83	27.95	57.15	41.21
en-ja	19.85	51.30	35.33	22.15	51.45	37.66	23.84	50.41	40.15	23.84	51.02	41.51
ja-en	19.25	37.47	27.81	20.89	38.37	29.7	20.77	41.87	30.98	22.01	40.37	32.20
en-fi	18.79	37.46	26.70	21.05	43.34	29.07	22.09	54.01	32.33	23.09	54.39	33.76
km-en	6.97	32.78	10.95	12.20	34.51	19.56	24.03	34.82	41.28	15.07	35.86	31.53
en-ta	5.38	29.87	11.28	7.09	38.93	16.44	12.82	51.57	28.81	12.36	42.88	28.22

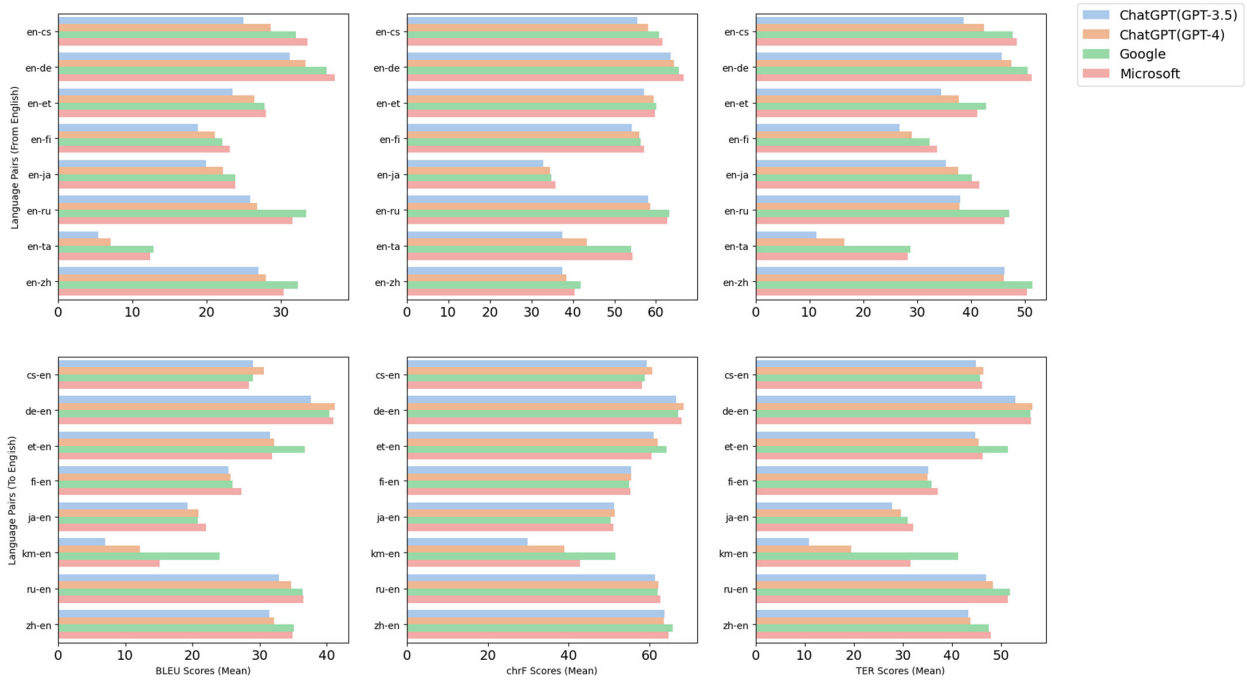

Figure 4. The comparison results by language pairs.

The chrF scores, which assess the quality of translations considering both the precision and recall of character sequences, reflect the pattern observed in the BLEU scores. Taking German–English (de-en) translation as an example, GPT-4 achieved a score of 68.43, followed by Microsoft Translator with 67.89, Google Translate with 67.06, and GPT-3.5 with 66.57. The TER scores of the translation systems continue to showcase a similar trend. For German–English (de-en) translation, GPT-4 took the lead with a score of 56.39, closely followed by Microsoft Translator with 56.16, Google Translate with 56.07, and GPT-3.5 with 52.93.

Examining the scores for diverse language pairs reveals a clear difference in performance among the translation systems. For example, when inspecting Russian–English (ru-en) translations, Google Translate takes the lead with a BLEU score of 36.41. Yet, in the same language pair, Microsoft Translator marginally surpasses others with a chrF score of 62.65, while Google Translate maintains its dominance with a TER score of 51.85. This observation suggests that a translation system's excellence in one metric does not necessarily secure its superiority in all metrics.

As we delve into lower-resource language pairs, the situation grows more complex. For example, in the translation from Khmer to English (km-en), Google Translate significantly outperforms others with a BLEU score of 24.03, a chrF score of 51.57, and a TER score of 41.28. In contrast, GPT-4 shows a diminished performance, attaining a BLEU score of 12.20, a chrF score of 38.93, and a TER score of 19.56. These differences underscore the challenges encountered and highlight the continuing need for advancements in machine translation for languages with fewer resources.

In conclusion, the performance of machine translation systems remains varied across different languages and evaluation metrics. While GPT-4 shows significant improvements over GPT-3.5 in many situations, other platforms, such as Google Translate and Microsoft Translator, continue to exhibit robust performance, especially with certain language pairs and metrics. These findings emphasize the ongoing need for development and refinement in the field of machine translation, with the aim of enhancing both the breadth and quality of translations across all language pairs.

4.5. The Impact of Reference Text Token Size

Table 7 presents the average BLEU, chrF, and TER scores grouped by the token size quartiles of the reference texts. The patterns observed indicate that the average token size in each quartile significantly impacts the performance of different models in varied ways. In the case of BLEU scores, an increase in the average token size corresponds to an improvement in scores across all models. This trend is demonstrated with Microsoft and Google Translate reaching a score of 32.18 when the average token size is at its largest (58.62), followed closely by GPT-4 with 30.75 and GPT-3.5 with 28.95.

Table 7. The results by quartiles of reference token size.

Section	Token Size (Mean)	ChatGPT (GPT-3.5)	ChatGPT (GPT-4)	Google	Microsoft
BLEU	10.59	22.90	25.35	27.89	27.97
	19.37	22.65	25.27	28.06	27.10
	29.14	24.70	26.15	29.82	28.90
	58.62	28.95	30.75	32.18	32.18
chrF	10.59	53.23	55.14	57.19	57.04
	19.37	52.24	54.90	57.05	56.40
	29.14	53.17	54.86	57.49	56.70
	58.62	55.75	57.26	57.80	57.90
TER	10.59	32.80	35.91	41.60	41.24
	19.37	33.72	36.77	42.60	41.31
	29.14	37.28	38.96	44.64	43.57
	58.62	43.48	45.00	47.17	47.25

In the case of TER scores, a pattern akin to the BLEU scores is observable, with scores escalating alongside an increase in the average token size. Upon reaching the largest average token size, 58.62, Microsoft Translator prevails with a score of 47.25, followed closely by Google Translate with 47.17, GPT-4 with 45.00, and GPT-3.5 with 43.48.

Meanwhile, the variation in chrF scores is relatively modest as the average token size increases. When the mean token size is at its largest at 58.62, the scores range between 55.75 for GPT-3.5 and 57.90 for Microsoft Translator. This implies that the impact of token size on character-level translation accuracy is considerably less.

According to the BLEU and TER metrics, these findings suggest that larger token sizes contribute to improved translation performance. However, the effect is less pronounced when considering the chrF metric. These observations underscore the intricate relationship between the characteristics of the reference text, such as token size, and the translation performance, thus highlighting the necessity for translation models to balance diverse factors.

Figure 5 illustrates the mean values of BLEU, chrF, and TER scores categorized by quartiles of the reference token size. Echoing the findings in the table, the graphical depiction illustrates an upward trend in BLEU and TER scores as the average token size increases. This trend strengthens the premise that texts with larger token sizes might provide more substantial context, thereby enhancing the performance of the translation models. However, the impact on chrF scores is less pronounced, emphasizing the unique ways in which different metrics respond to variations in token size.

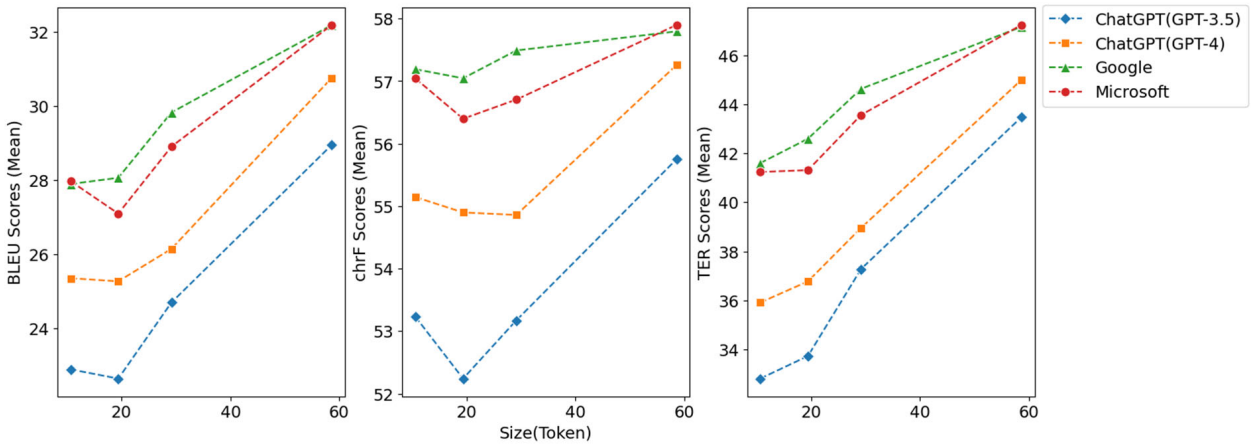

Figure 5. The score comparison by quartiles of reference token size.

5. Conclusions

5.1. Findings and Discussion

This study analyzed the translation performances of LLMs and NMT systems, considering various factors such as the systems used, languages, and reference token size. Among the key findings, we observed that for translations from English to non-English, Google Translate and Microsoft Translator demonstrated superior performance, while GPT-3.5 and GPT-4 exhibited comparatively less robust results. These insights indicate the ongoing competitiveness of traditional translation services such as Google Translate and Microsoft Translator when dealing with translations from English to other languages.

In summary, across BLEU, chrF, and TER scores, Google Translate consistently outperforms ChatGPT and Microsoft Translator, highlighting its overall superior translation quality in the assessed aspects. The performance improvement observed between GPT-3.5 and GPT-4 in translation quality can be attributed to advancements in the underlying architecture, training data, and optimization techniques. GPT-4's increased capacity enables it to handle more complex language structures and provide more accurate translations. As ChatGPT continues to evolve, incorporating novel approaches and leveraging state-of-the-art research, it is anticipated that further enhancements may bring its performance closer to, or surpass, that of other leading translation services such as Google Translate and Microsoft Translator.

Furthermore, the performance gap between the translation systems is more substantial for certain language pairs, such as English–Tamil (en-ta) and Khmer–English (km-en), suggesting potential areas for improvement in these specific translation tasks. Conversely, for language pairs such as Czech–English (cs-en), German–English (de-en), and Estonian–English (et-en), the performances of the translation systems are relatively close, implying that these systems are more evenly matched for these translation tasks. Interestingly, GPT-4 surpasses the performance of both Google and Microsoft in specific language pairs. For instance, in the German–English (de-en) language pair, GPT-4 achieves higher scores in all metrics. This notable performance improvement suggests promising potential for future enhancements of the GPT model, even in challenging translation tasks.

While the BLEU, chrF, and TER scores provide valuable insights into the translation qualities of these systems, they are not the sole indicators of translation performance. Other factors, such as fluency, adequacy, and the preservation of meaning, should also be considered when evaluating translation systems. Despite this, these results offer a valuable foundation for understanding and comparing the performance of GPT-3.5, GPT-4, Google Translate, and Microsoft Translator across various language pairs and highlight the promising potential of next-generation GPT models.

In addition, the results indicate a potential boundary condition pertaining to the evaluated NMT systems and LLMs, including Google Translate, Microsoft Translator,

and ChatGPT. These systems may have been previously trained on WMT data or similar datasets. If this is the case, our experimental approach, which heavily relies on WMT datasets, might have introduced certain biases, which could influence the broad applicability of our conclusions when these systems encounter novel linguistic contexts. Future research should consider this factor and incorporate more diverse datasets to ensure a more comprehensive and robust evaluation of both NMT systems and LLMs.

Lastly, the data indicated a general enhancement in the performance of machine translation models as the token size in the reference texts increased. Particularly, reference texts with the largest token sizes led to the highest performances across all three models in terms of BLEU, chrF, and TER scores. Conversely, reference texts with the smallest token sizes resulted in the lowest performances. This suggests that translation models may excel more when processing texts with larger tokens, as they can capture more contextual nuances and produce translations that are more closely aligned with the reference translations in terms of lexical similarity, character-level precision and recall, and overall translation quality.

Additionally, this finding suggests larger tokens rich in unique details enable models to generate more precise translations. Conversely, fewer tokens may limit context, thus challenging translation quality. These results highlight token size's role in translation performance, suggesting that future research focuses on improving the handling of smaller tokens and expanding their context to enhance accuracy across varying token sizes. In conclusion, these findings contribute to our understanding of the translation performances of LLMs and NMT systems, providing insights for researchers and developers to further enhance translation quality and explore the potential of next-generation models.

5.2. Research Implications

This study highlights the importance of conducting direct comparisons between the performances of LLMs and machine translation systems from a user's perspective and emphasizes the potential of utilizing LLMs as translation tools. Based on these research findings, several practical implications can be suggested.

The performance of LLMs' translation capabilities varied across different language pairs, with some pairs demonstrating larger performance gaps than others. This finding highlights the need for researchers and developers to focus on enhancing translation models for underperforming language pairs. The analysis results indicate that longer reference texts positively influence machine translation performance, implying that incorporating longer reference texts during the training process could result in better-performing translation models. Building on these observations, future efforts should emphasize the following areas.

First, one approach is to enhance the fine-tuning process for underperforming language pairs with a diverse and comprehensive dataset, which could considerably improve their performance. Prioritizing the inclusion of longer reference texts during training is another important step. This strategy could heighten a model's proficiency in complex translations and significantly improve the overall translation performance. Such concerted efforts may guide the ongoing advancement of LLMs toward more accurate and efficient translation systems.

Second, since ChatGPT is primarily designed as a conversational AI model, its translation performance might not match specialized translation systems. However, its potential for more natural, conversation-like interactions could still be valuable in specific use cases. In this regard, one promising avenue could be to leverage ChatGPT's conversational capabilities in areas such as real-time customer service or community management on social platforms. In these instances, the model's ability to deliver more organic and interactive translations could significantly enhance the quality and fluidity of cross-language communications, thereby fostering more engaging and meaningful interactions. This exemplifies a practical scenario where the unique strengths of ChatGPT could be effectively harnessed, including in the NMT field.

Third, investigating domain-specific translation performance is encouraged. Translation systems might perform differently when handling domain-specific content, such as scientific articles, legal documents, or news articles. Future research could examine how the performance of machine translation systems varies across different domains. Additionally, exploring the impact of data scarcity and low-resource languages on translation performance is suggested. The performance of LLMs might be significantly influenced by the availability of training data for particular languages.

To address this challenge, companies should expand their training data to specific domains and low-resource languages. This could be achieved using techniques such as web scraping, the collection of specialized documents, or crowdsourcing. Furthermore, developing new algorithms and technologies that aid in understanding the specific requirements and context of a given domain in collaboration with domain experts is crucial for enhancing translation quality for specific domains and low-resource languages. Thus, the usefulness and diversity of machine translation systems could be improved, leading to an enhanced user experience.

5.3. Research Limitations and Future Plans

Despite the significance and implications of this research, this study has the following limitations. First, our research focused on a select set of language pairs, potentially not reflecting the performance of translation systems across all languages. It also has a weak data size of 200 sentence pairs per language pair. Therefore, future research could expand this analysis to include a more diverse range of languages and language pairs to obtain a comprehensive understanding of translation systems' performance.

Second, this study primarily relied on BLEU, chrF, and TER scores to assess translation quality. While these metrics are widely used and standard, they may not capture all aspects of translation performance. Notably, this study did not include state-of-the-art metrics such as COMET and BLEURT, which utilize trained models for evaluation, due to potential bias concerns in providing an objective performance evaluation. Future research, particularly studies building on transformer models, may consider these metrics. Moreover, incorporating human evaluations or alternative translation quality assessment methods would give a more nuanced understanding of translation system performance. In addition, the exploration of other factors, such as fluency, adequacy, and preservation of meaning, could provide a more comprehensive evaluation of machine translation systems.

Third, this study used a single evaluation dataset from 2018–2020, which may not reflect the dynamic nature of real-world translation tasks. Investigating the performance of translation systems using multiple evaluation datasets and covering various domains, genres, and text types could provide better insights into how these systems perform in different contexts.

Fourth, this study focused on commercially available translation models without including open-source models. This requires future performance exploration of NMT models such as Opus and decoder-only LLMs such as Bloom and Llama. In addition, research that can draw statistical conclusions that verify researchers' differentiated hypotheses or research problems based on the open-source model should be conducted to verify effectiveness or influence beyond simple trends or comparisons.

Lastly, this study assessed ChatGPT's performance as a representative of LLMs. With the continuous fine-tuning and advancement of machine learning models, ChatGPT is expected to undergo regular updates, and new LLMs may emerge that potentially surpass ChatGPT or offer alternative language model options. Future research should evaluate the performance of updated ChatGPT models and emerging LLMs to track their progress, identify trends in their development, and enable more comprehensive comparisons between different models.

Author Contributions: Conceptualization, J.S. and B.K.; methodology, J.S.; software, J.S.; validation, J.S.; formal analysis, J.S.; investigation, J.S.; resources, J.S.; data curation, J.S.; writing—original draft preparation, J.S. and B.K.; writing—review and editing, B.K.; visualization, B.K.; supervision, B.K.; project administration, B.K.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 121–154. [CrossRef]
2. Acumen Research and Consulting. Available online: https://www.acumenresearchandconsulting.com/press-releases/machine-translation-market (accessed on 20 May 2023).
3. Biswas, S.S. Potential use of chatGPT in global warming. *Ann. Biomed. Eng.* **2023**, *51*, 1126–1127. [CrossRef] [PubMed]
4. Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.L.; Tang, Y. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1122–1136. [CrossRef]
5. Mathew, A. Is artificial intelligence a world changer? A case study of OpenAI's Chat GPT. *Recent Prog. Sci. Technol.* **2023**, *5*, 35–42.
6. Meng, F.; Zhang, J. DTMT: A novel deep transition architecture for neural machine translation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 224–231.
7. Hutchins, J. The history of machine translation in a nutshell. *Retrieved Dec.* **2005**, *20*, 1–5.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 1–11.
9. Chand, S. Empirical survey of machine translation tools. In Proceedings of the 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 23–25 September 2016; IEEE: New York, NY, USA, 2016; pp. 181–185.
10. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 18 April 2023).
11. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
12. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 3–10 March 2021; pp. 610–623.
13. Hutchins, W.J.; Somers, H.L. *An Introduction to Machine Translation*; Academic Press Limited: London, UK, 1992.
14. Wang, H.; Wu, H.; He, Z.; Huang, L.; Church, K.W. Progress in Machine Translation. *Engineering* **2022**, *18*, 143–153. [CrossRef]
15. Taravella, A.; Villeneuve, A.O. Acknowledging the needs of computer-assisted translation tools users: The human perspective in human-machine translation. *J. Spec. Transl.* **2013**, *19*, 62–74.
16. Rodríguez-Castro, M. An integrated curricular design for computer-assisted translation tools: Developing technical expertise. *Interpret. Transl. Train.* **2018**, *12*, 355–374. [CrossRef]
17. Ragni, V.; Nunes Vieira, L. What has changed with neural machine translation? A critical review of human factors. *Perspectives* **2022**, *30*, 137–158. [CrossRef]
18. Chopra, D.; Joshi, N.; Mathur, I. Improving translation quality by using ensemble approach. *Eng. Technol. Appl. Sci. Res.* **2018**, *8*, 3512–3514. [CrossRef]
19. Hearne, M.; Way, A. Statistical machine translation: A guide for linguists and translators. *Lang. Linguist. Compass* **2011**, *5*, 205–226. [CrossRef]
20. Hutchins, J. Example-based machine translation: A review and commentary. *Mach. Transl.* **2005**, *19*, 197–211. [CrossRef]
21. Cui, Y.; Surpur, C.; Ahmad, S.; Hawkins, J. A comparative study of HTM and other neural network models for online sequence learning with streaming data. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; IEEE: New York, NY, USA, 2016; pp. 1530–1538.
22. Mara, M. *English-Wolaytta Machine Translation Using Statistical Approach*; St. Mary's University: San Antonio, TX, USA, 2018.
23. Maruf, S.; Saleh, F.; Haffari, G. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–36. [CrossRef]
24. Zhu, J.; Xia, Y.; Wu, L.; He, D.; Qin, T.; Zhou, W.; Li, H.; Liu, T. Incorporating BERT into Neural Machine Translation. *arXiv* **2020**, arXiv:2002.06823.
25. Kulshreshtha, S.; Redondo-García, J.L.; Chang, C.Y. Cross-lingual alignment methods for multilingual BERT: A comparative study. *arXiv* **2020**, arXiv:2009.14304.

26. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. *Assoc. Comput. Linguist.* **2020**, 8440–8451. [CrossRef]
27. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
28. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
29. OpenAI. GPT-4 Is OpenAI's Most Advanced System, Producing Safer and More Useful Responses. Available online: https://openai.com/gpt-4 (accessed on 28 May 2023).
30. Freedman, J.D.; Nappier, I.A. GPT-4 to GPT-3.5:'Hold My Scalpel'—A Look at the Competency of OpenAI's GPT on the Plastic Surgery In-Service Training Exam. *arXiv* **2023**, arXiv:2304.01503.
31. Koehn, P.; Haddow, B. Interactive assistance to human translators using statistical machine translation methods. In Proceedings of the Machine Translation Summit XII: Papers, Ottawa, ON, Canada, 26–30 August 2009.
32. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
33. Callison-Burch, C.; Osborne, M.; Koehn, P. Re-evaluating the role of BLEU in machine translation research. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006; pp. 249–256.
34. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; pp. 392–395.
35. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
36. Wieting, J.; Berg-Kirkpatrick, T.; Gimpel, K.; Neubig, G. Beyond BLEU: Training Neural Machine Translation with Semantic Similarity. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4344–4355.
37. Castilho, S.; Moorkens, J.; Gaspari, F.; Calixto, I.; Tinsley, J.; Way, A. Is neural machine translation the new state of the art? *Prague Bull. Math. Linguist.* **2017**, *108*, 109–120. [CrossRef]
38. Callison-Burch, C.; Koehn, P.; Monz, C.; Peterson, K.; Przybocki, M.; Zaidan, O. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, 15–16 July 2010; pp. 17–53.
39. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
40. Nichols, J.; Warnow, T. Tutorial on computational linguistic phylogeny. *Lang. Linguist. Compass* **2008**, *2*, 760–820. [CrossRef]
41. Birch, A. *Neural Machine Translation*; Cambridge University Press: Cambridge, UK, 2021.
42. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.G.; Le, Q.; Salakhutdinov, R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2978–2988.
43. Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* **2023**, *15*, 192. [CrossRef]
44. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 1–9.
45. Hariri, W. Unlocking the Potential of ChatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. *arXiv* **2023**, arXiv:2304.02017.
46. Post, M. A call for clarity in reporting BLEU scores. *arXiv* **2018**, arXiv:1804.08771.
47. WMT 18. Available online: https://www.statmt.org/wmt18/translation-task.html (accessed on 15 April 2023).
48. WMT 20. Available online: https://www.statmt.org/wmt20/translation-task.html (accessed on 15 April 2023).
49. Koehn, P.; Chaudhary, V.; El-Kishky, A.; Goyal, N.; Chen, P.J.; Guzmán, F. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 19–20 November 2020; pp. 726–742.
50. Bojar, O.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Monz, C. *WMT18*; Association for Computational Linguistics: Belgium, Brussels, 2018.