https://doi.org/10.31470/2309-1797-2024-35-2-134-156 UDC 81'23

Artificial Intelligence in Academic Translation: A Comparative Study of Large Language Models and Google Translate [*]

Штучний інтелект в академічному перекладі: Порівняльне дослідження великих мовних моделей та Google Translate [**]

Mohammed Ali Mohsen
Ph.D. in Linguistics,
Professor of Applied linguistics

Мохаммед Алі Мохсен
доктор філософії в галузі
лінгвістики, професор у галузі
прикладної лінгвістики

E-mail: mmohsen1976@gmail.com
mamohsen@nu.edu.sa
https://orcid.org/0000-0003-3169-102X

Najran University
(Saudi Arabia)
✉ King Abdulaziz Street,
Najran, Kingdom of Saudi Arabia,
66251

Університет Наджран
(Саудівська Аравія)
✉ вул. Короля Абдулазиза,
Наджран, Саудівська Аравія,
66251

Original manuscript received December 15, 2023
Revised manuscript accepted March 18, 2024

[*] This research received grant no. (15/2023) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

[**] Це дослідження отримало грант № (15/2023) від Арабської обсерваторії перекладу (філії ALECSO), яку підтримує Комісія з питань літератури, видавничої справи та перекладу Саудівської Аравії.

ABSTRACTS

Purpose. *The advent of Large Language Model (LLM), a generative artificial intelligence (AI) model, in November 2022 has had a profound impact on various domains, including the field of translation studies. This motivated this study to conduct a rigorous evaluation of the effectiveness and precision of machine translation, represented by Google Translate (GT), in comparison to Large Language Models (LLMs), specifically ChatGPT 3.5 and 4, when translating academic abstracts bidirectionally between English and Arabic.*

Methods. *Employing a mixed-design approach, this study utilizes a corpus comprising 20 abstracts sourced from peer-reviewed journals indexed in the Clarivate Web of Science, specifically the Journal of Arabic Literature and Al-Istihlal Journal. The abstracts are equally divided to represent both English-Arabic and Arabic-English translation directionality. The study's design is rooted in a comprehensive evaluation rubric adapted from Hurtado Albir and Taylor (2015), focusing on semantic integrity, syntactic coherence, and technical adequacy. Three independent raters carried out assessments of the translation outputs generated by both GT and LLM models.*

Results. *Results from quantitative and qualitative analyses indicated that LLM tools significantly outperformed MT outputs in both Arabic and English translation directions. Additionally, ChatGPT 4 demonstrated a significant advantage over ChatGPT 3.5 in Arabic-English translation, while no statistically significant difference was observed in the English-Arabic translation directionality. Qualitative analysis findings indicated that AI tools exhibited the capacity to comprehend contextual nuances, recognize city names, and adapt to the target language's style. Conversely, GT displayed limitations in handling specific contextual aspects and often provided literal translations for certain terms.*

Key words: *ChatGPT, Machine Translation, Google Translate, articles' abstract.*

Introduction

The rapid advancement of technology has had a profound impact on various aspects of human life, including the field of translation studies. For many years, researchers, as highlighted by Quah (2006), have been striving to automate the translation process, aiming to convert text from a source language (SL) to a target language (TL). However, this endeavor has been challenged by the complexity of human language, which includes nuances and idiomatic expressions unique to each language. These challenges have led to doubts regarding the effectiveness of machine translation (MT). In fact, historical records, such as the PAALC report from the 1960s, cautioned against the use of MT due to its limitations in accurately conveying intended meanings. The emergence of Statistical Machine Translation (SMT) in the late 1980s rekindled optimism in the field, suggesting that while MT might not achieve perfect accuracy, it could still serve as a useful tool requiring human intervention for quality assurance. The landscape of MT underwent a transformative change with the introduction of Neural Machine Translation (NMT) by Google in 2016. Utilizing advanced natural language processing techniques, NMT significantly enhanced the quality of translated text. This paradigm shift led to a transition from a Computer-Assisted Translation (CAT) model to a Machine-Assisted Human Translation (MAHT) model, positioning technology as the principal agent in the translation process (Mohsen et al., 2023). In recent years, the advent of Large Language Model (LLM), particularly the Generative Pre-trained Transformer (ChatGPT) developed by OpenAI, has further revolutionized the field (Baidoo-Anu & Ansah, 2023; Lyu et al., 2023). Initially designed as a conversational interface, ChatGPT has demonstrated a wide array of functionalities beyond dialogic interactions, including translation capabilities (Hendy et al., 2023). The launch of the new AI tool has motivated the researchers to investigate its efficacy in terms of translating academic genre. Given the potentials of AI and Google translate to automate translating texts and the variability across type of language and text genre, the current study aims to evaluate their efficacy in aiding the English to Arabic translation process. By identifying their strengths and limitations, this study offers valuable insights to translators on optimizing their workflow. Furthermore, pinpointing areas for improvement in AI and MT will not only enhance their utility but also contribute to the broader discourse in translation studies. This investigation thus serves a dual purpose: facilitating immediate practical applications and informing future technological advancements in the field. Therefore, the current study aims to address the following research question:

RQ 1. How do the translational outputs of ChatGPT and Google Translate compare in terms of good quality translated outputs when translating academic abstracts between Arabic and English?

Literature Review

AI Affordances and Limitations

Over the past several decades, Artificial Intelligence (AI) has undergone transformative developments, culminating in sophisticated algorithms that address intricate problems and streamline human endeavors (Hossain, 2023). A salient manifestation of this progress lies in the realm of MT, which utilizes advanced computational models to transcribe textual or auditory content from one language to another (Sennrich et al., 2017; Vaswani et al., 2018). This sector has particularly benefited from the infusion of deep learning paradigms (Goodfellow et al., 2016). Fundamentally, AI architectural design is anchored in the seminal Transformer model by (Vaswani et al., 2018), a blueprint that has been pivotal in shaping contemporary MT technologies. This is exemplified by Google's Transformer-based Neural Machine Translation system (GNMT), which attests to the model's scalability and efficiency (Wu et al., 2018). In its operational schema, ChatGPT adopts a comprehensive data-driven methodology, assimilating a diverse corpus spanning numerous linguistic styles and thematic domains (Malik et al., 2023). While this renders ChatGPT a versatile instrument for translation endeavors, it is imperative to acknowledge that ChatGPT lacks adequacy in providing a good text outputs for low-sources languages and sometimes generate subtle word-level hallucinations (Hendy et al., 2023). This issue motivates the researcher to investigate how different MT tools treated Arabic academic texts, being a low-sourced language (Islam et al., 2021) to English, and in return, how English texts would be translated into Arabic.

MT systems like Google Translate, which has been thoroughly investigated in translation studies literature, have demonstrated efficacy in real-time text-based interactions and broad language coverage (Jiao et al., 2023). Recent research indicates that chatbot language models like ChatGPT offer unique affordances that warrant further investigation. Specifically, ChatGPT has shown competitive performance in translating high-resource European languages, albeit with limitations in low-resource languages (Jiao et al., 2023). Beyond mere translation, chatbots have been found to enhance user experience in various industries such as travel, tourism, and hospitality (Bulchand-Gidumal et al.). They operate through a multi-step process involving natural language understanding, automatic response generation, and fluent language construction, offering the potential for more nuanced and context-aware translations (Suta et al., 2020). Additionally, chatbots have demonstrated cost and time efficiency in translational medicine (Abashev et al., 2017) and have even outperformed teacher counseling services in educational settings (Wu et al., 2020). These unique affordances suggest that chatbot language models could complement or even enhance existing MT technologies, making them a subject of significant academic interest.

While MT and AI can help human translators to speed up the translation process, it is critical to acknowledge that MT is not asolution for addressing the multifaceted nature of natural language. The complexity of human language, which is influenced by factors such as text genre, contextual nuance, and the speaker or writer's intent, poses considerable challenges for the accurate rendering of meaning through MT alone (Valdeón, 2023). As such, it is often imperative to involve human expertise in various MT contexts. Human intervention becomes indispensable for tasks like disambiguation, interpretation of idiomatic expressions, and ensuring cultural sensitivity, among other complex linguistic and semantic aspects that MT systems might not fully grasp (Hutchins, 2005). This underlines the necessity for a hybrid model that synergizes both machine and human capabilities to achieve higher levels of accuracy and contextual relevance. Therefore, it is crucial to not overstate the capabilities of MT and to recognize the continuing need for human expertise in ensuring the fidelity and nuanced understanding of translated text.

Literary Text and Translation

The literary genre is known for evoking emotion rather than providing information, which requires additional reading and listening to be fully comprehensible (Jones, 2019). Some authors leave the interpretation of their texts up to the readers, allowing them to understand it in their own way. This can result in multiple meanings and make translation a daunting task (Ponzio, 2007). Unlike technical translation, literary translation is challenging because of the multiple meanings the text contains and the various intentions that the narrators, novelists, and poets want to convey. As a result, it was previously believed that this genre could only be handled by human translators and that any attempt to automate literary translation with MT would be futile. However, the development of NMT has brought hope that machine can assist human translators in dealing with literary texts, although their reliability is not yet perfect (Sanz-Valdivieso & López-Arroyo, 2023). Literary translation presents unique difficulties that make it harder to automate compared to technical translations. According to Toral and Way (2015), literary texts often require more liberal or metaphorical translations, which can confuse alignment algorithms (Voigt & Jurafsky, 2012). Additionally, the broader range of topics and richer vocabulary found in the literary texts leads to texts that are less predictable for machine translation systems (Toral & Way, 2015). Other factors like preserving rhyme, rhythm, meter and other literary constraints, especially in poetry translation (Genzel et al., 2010), as well as managing dense discourse features like referential cohesion (Voigt & Jurafsky, 2012) further complicate the process. In the present study, the author seeks to investigate the efficacy of translating abstracts of literary academic texts employing two distinct approaches: LLMs, exemplified by ChatGPT, and MT, exemplified by GT. The primary objective is to assess the effectiveness of these methods and to gauge their ability to approximate the translation quality achieved by human translators.

Related Works

With the introduction of SMT and NMT, various attempts have been made to investigate the effectiveness and accuracy of MT in translating texts, as well as its potential to assist or even replace human translators. The nature of the text is crucial in determining whether MT can be a suitable tool for human translators, as some texts are easier to translate while others pose more challenges. The main question here is whether MT can successfully translate literary texts in different or related languages, which remains a challenge for researchers.

In an effort to address this question, Toral and Way (2015) conducted a study exploring the use of SMT for literary translation between two related languages: Spanish and Catalan. Their aim was to assess the translatability of literary texts and evaluate the feasibility of using SMT for translating novels. The authors utilized corpus analysis to compare the domain specificity and translation freedom of literary parallel texts with other domains. The findings revealed that while translating novels posed more constraints compared to news texts, it was less restrictive than translating technical texts. To translate a novel, the authors adapted an ES-CA SMT system and utilized in-domain novels for tuning, language modelling, and translation modelling. As a result, the adapted SMT system showed an improvement of +9.38% Bilingual Evaluation Understudy (BLU) over a strong baseline. Further analysis showed that nearly 20% of the sentences had matches at the sentence level with the human translation, with an additional 10% requiring only small edits (within 5 edits). A manual evaluation indicated that over 60% of the MT output was rated as good as or even better than the professional translation. Toral and Way (2015) concludes that while SMT can assist in literary translation for related languages, challenges at the discourse level still need to be addressed.

As the launch of ChatGPT was in November 2022, there is a dearth of research studies investigating the efficacy of such tool on improvement of translation studies. Several works appeared online as pre-prints that are yet published due to the long peer review process. Two preprints are related to the use of ChatGPT to the field of translation studies in low-sourced languages context: Arabic and Bengali. The first attempt was conducted by Ghosh and Caliskan (2023) who examined gender bias in ChatGPT when translating between English and Bengali, a widely spoken yet understudied language that uses gender-neutral pronouns. Through prompts about occupations and actions, the authors find ChatGPT exhibits strong implicit gender biases, associating certain occupations like doctor with ‘he’ and nurse with 'she'. It also associates certain actions like cooking with ‘she’. ChatGPT fails to translate the English gender-neutral pronoun ‘they’ properly into Bengali. The biases persist even when gender information is explicitly provided in the prompt. Similar issues occur when translating from five other gender-neutral languages into English. The authors situate the biases as stemming from the training data and socio-technical factors privileging high-resource languages like English. They argue the biases can perpetuate harmful gender stereotypes and erase non-binary identities. The authors call for a human-centered approach to designing translation tools that meaningfully involves speakers of diverse languages.

The second attempt was carried out by Khoshafah (2023) who evaluates the accuracy of ChatGPT for Arabic-English translation by comparing its outputs to professional human translations across various text genres like historical, literary, media, legal, and scientific. The results show ChatGPT can provide accurate translations for simple content but struggles with complex texts requiring domain knowledge or cultural nuance. Though generally conveying the right meaning, ChatGPT's translations lack the precision of human translations for legal documents, medical reports, scientific studies, and literary works. The author concludes that ChatGPT is a valuable tool for basic cross-cultural communication, but limitations remain compared to human translation. They offer recommendations for using ChatGPT as a translator, emphasizing the need for caution with technical, culturally specific, or highly sensitive content, and combining it with professional translation/ proofreading where accuracy is critical.

While the previous attempts have made progress in advancing the field by exploring how LLM tools can be used for translating low-sourced languages like Arabic and Bengali, it is crucial to thoroughly investigate this in comparison to the dominant MT tool, GT. Our focus is to examine abstracts of academic literary texts due to their succinct yet information-rich composition and to analyze how different tools handle various language pair directions, such as Arabic-English and English-Arabic.

Methodology

Situated within a mixed-methods research paradigm, the current study was explicitly designed to scrutinize the translation of academic abstracts between the English and Arabic. Given the inherent complexities and specialized lexicon endemic to scholarly discourse, the investigation aims to rigorously assess the fidelity and accuracy of translations produced by GT and ChatGPT. The empirical corpus for this inquiry encompasses a total of 20 research article abstracts, equally partitioned between the translational directions of English-Arabic and Arabic-English. These abstracts are elicited from two scholarly peer-reviewed journals in Arabic literature discourse, namely, *Journal of Arabic Literature* and *Al-Istihlal Journal.* Four issues were the rationale for selecting these specific journals as the corpus source. Firstly, the two journals' main scope is to publish research examining Arabic literary works, ensuring the originality of the SL texts and close approximation to the areas being investigated. Secondly, the two journals are distinguished for their international scope and adopting rigorous peer-review process, thereby enhancing scholarly credibility. Lastly, their indexing in the Clarivate Web of Science Core Collection serves as an additional endorsement of the academic rigor embedded in its published material. Fourth, the abstracts are available in the Web of Science and could be downloaded to an excel sheet that to be easily scrutinized.

Data Extraction

The titles of the journals were inputted into the Web of Science database to retrieve abstracts of research papers pertaining to literary works associated with the Arabic language. Our focus was limited to journals dedicated exclusively to the examination of Arabic literature. Among these journals, we specifically chose the *Journal of Arabic Literature*, published by Brill, which features scholarly articles in English. Additionally, we selected the *Al-Istihlal journal*, which publishes articles in Arabic, accompanied by abstracts available in both Arabic and English. These texts are generally dense with discipline-specific terminology and employ a high degree of academic formality, which can serve as a reliable benchmark for assessment and might pose a challenge for both tools. Comprehensive records, including the article titles, authors, publication sources, and abstracts, were extracted and downloaded in the form of an Excel spreadsheet. Regardless of their publication dates or the topics covered, I randomly selected 10 abstracts for further analysis and examination.

Abstracts Translation

In order to streamline the translation process for the selected abstracts, I implemented the GPT for Excel Word tool, integrating Chat GPT 3.5-turbo for efficient automated translation. Within Excel, we employed the function “GPT_translate (source text, ‘target text’)” to facilitate bidirectional translation between Arabic and English for the two files at hand. Moreover, I leveraged Google Sheets to harness the functionalities associated with GT, utilizing the specific function “GOOGLE_translate (source language, target language)”. Concerning the use of ChatGPT 4, the abstracts were placed in the search bar one by one using the following prompt (Please translate the following abstract to English/Arabic) as to keep giving the ChatGPT 4 the translation outputs. The resulting translations, including those generated by the LLM, and those from GT were subsequently transcribed into the corresponding Excel sheets. This approach enabled us to consolidate and juxtapose the translated versions alongside the original content for comprehensive analysis and comparison.

Rubric and Evaluation

To evaluate the outputs performed by LLM and GT, the translated texts were scrutinized to evaluative rubric formulated by (Hurtado Albir & Taylor, 2015), which emphasizes key dimensions such as adequacy of the SL meaning, syntactic coherence, technical adequacy, and correct use of vocabulary. Three main categories were the focus of the rubric, namely, expression of the meaning of the original text (40%), which contains three subcategories; the same information, the same clarity, and the same register. The second category entitles "Composition in the target language (40%), encompassing four subcategories, Conventions of written language (correct orthography and typography, Vocabulary (appropriateness and richness, Morphosyntax (good use of verb tenses and modes, prepositions, etc.), Cohesion (good use of connectors and referential elements), and Coherence (ideas well organized and clearly presented). Details are presented in the screenshot (Appendix A). All these categories were placed near the texts and scores were shown for raters for the purpose of assessment. Three raters were recruited to assess the outputs of the translated texts against the rubric. Those raters have been working as instructors of translation courses, one is an Arabic native speaker who is working as an assistant professor of English literature, another one is MA holder of translation studies working as a lecturer of translation studies, and a third one is an MA holder of applied linguistics and an expert in language assessment. To avoid hallo effect, texts were made anonymous in terms of the translation tools and instead they were coded as evaluation 1, evaluation 2, and evaluation 3 (referring to GT, ChatGPt 3.5, and ChatGPT 4 respectively). The raters evaluated each translated abstracts based on the established rubric, and their scores were combined and averaged to run statistical analysis in SPSS software to identify the performance scores attributable to each tool. Subsequently, a qualitative examination of the types of errors and inadequacies are conducted, offering insights into the respective strengths and limitations of ChatGPT 3.5, 4 and GT.

Data Analysis

This study sought to evaluate the accuracy of translations produced by three different methods: GT, ChatGPT 3.5 and ChatGPT 4.

The analysis involved both descriptive data that calculate the means and standard deviation for all types of translated texts. Additionally, inferential statistical procedures were run to find out the comparisons between every type of translation. To run the data set of the current study, normal distribution of the data was run using A Shapiro-Wilk test considering that the study sample was small. Consequently, a non-parametric Kruskal-Wallis test was conducted to determine the performance of three scenarios investigated in the study against the accuracy of the translation outputs. The scores were averaged based on the raters' evaluation and placed in the SPSS file. I use Paython to highlight matching words that appear in every translation tool for the two genres and calculated the rate of matching words among the tools in general and between the two AI tools in specific (See supplementary materials B and C).

Results

Quantitative Analysis

English-Arabic Text Directionality

The initial analysis provided for the translation scores, as presented in Table 1. ChatGPT 4 (GPT4) recorded the highest average score (Med = 9.35, SD = 0.84), followed by ChatGPT 3.5 (GPT3.5) (M = 8.9, SD = 0.54), and GT had the lowest average score (Med = 4.75, SD = 0.43). This suggests a preliminary indication of the superiority of ChatGPT 4 in translation accuracy.

***Table* 1**
Descriptive and Inferential Statistics for the English Abstracts Translated into Arabic

Translation tool	Median	Mean rank	SE	H	Df	P	r(effect size)
Google Translate	4.75	5.50	.26	21.14	2	.000	0.92
ChatGPT 3.5	8.9	16.00	.17				
ChatGPT 4	9.35	25.00	.13				

Table 1 shows that there was a trajectory improvement from GT to ChatGPT 3.5 to ChatGPT 4. To see if the scores across tools were statistically significant, a post-hoc comparison was run and summarized in Table 2.

***Table* 2**
A Post-hoc Comparison of Scores of Translation Tools for English-Arabic Directionality

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Effect size (r)
ChatGPT3.5-GT	10.5	3.9	2.66	.023*	.84
ChatGPT4-GT	19.5	3.9	4.95	.000*	.86
ChatGPT3.5-ChatGPT 4	9.0	3.9	2.28	.067	-

Further analysis using Kruscal Wallis with Bonferroni correction for pairwise comparisons indicated significant differences between each pair of translation methods (see Table 2). Results showed that both LLM translation tools significantly outperformed GT, *H* (2)=21.1, *p*<.05, *r*= 92. Specifically, ChatGPT 3.5 statistically outscored English texts translated into Arabic, as indicated by Mann-Whiteney Test *U*(N=10)= .000, *z*=3.72, *p*<.05, *r*=.84. Similarly, ChatGPT 4 scored significantly better that GT, *U*(N=10)= .000, *z*=3.72 *p*<.05, *r*=.86. However, there was no significant difference that was reported between the scores of texts translated by ChatGPT 3.5 or ChatGPT 4.

English-Arabic Directionality

Another descriptive and inferential statistics were run to calculate the median scores of the Arabic literary texts translated to English and to see if translation tools outperform each other. Table 3 summarizes that.

***Table* 3**
Descriptive and Inferential Statistics for the English Abstracts Translated into English

Translation tool	Median	Mean Rank	SE	H	Df	P	r(effect size)
Google Translate	4.52	5.50	3.93	21.4	2	.000	0.95
ChatGPT 3.5	8.24	17.90	3.93				
ChatGPT 4	9.67	23.10	3.93				

A non-parametric test, Kruskal Wallis was utilized to run the statistical analysis across the texts scores. The difference between the median along with the mean ranks were significant, H(2)=24.6, p<.05, r=.95. Post hoc comparisons were conducted using Mann-Whitney Tests with a Bonferroni adjusted alpha correction. The difference between Google Translate and ChatGPT 3.5 was statistically significant (*U* (*N=10*) = .00, *z* = 3.71, *p* = .000, *r*=84). Concerning the comparison between the performance of GT versus ChatGPT 4, results from the Mann-Whitney Test showed significant differences, *U*(N=10)= .000, *z*=3.79, *r*=90. Likewise, results indicated a superiority of ChatGPT 4 performance over ChatGPT 3.5 (*U*=10), .000, *z* = 3.78, *p* = .000, r=.75). Again, a post-hoc analysis was performed to see statistical significance over the three translation tools for Arabic-English directionality.

***Table* 4**
A post-hoc Comparison of Scores of Translation Tools for Arabic-English Directionality

Sample 1-Sample 2	**Test Statistic**	**Std. Error**	**Std. Test Statistic**	**Sig.**	**Effect size (r)**
ChatGPT3.5-GT	12.9	3.9	3.15	.000*	.84
ChatGPT4-GT	17.6	3.9	4.47	. 000*	.90
ChatGPT3.5-ChatGPT 4	5.20	3.9	1.32	.002*	.75

The results from Table 4 underscore significant differences in the quality of translations provided by Google Translate, ChatGPT 3.5, and ChatGPT 4. The advanced AI models, particularly ChatGPT 4, exhibited a notable superiority in translation accuracy over traditional machine translation tools like Google Translate. These findings not only highlight the rapid advancements in AI-driven language translation but also emphasize the increasing effectiveness of these models in producing translations that are closer to human-level quality**.**

Qualitative Analysis

To enhance the findings from qualitative analysis, I analysed one abstract from each genre to find out how translation ran over every translation tool against accuracy and readability of the translation outputs in the target language. Given the constraints of word counts of this article, I will critically examine one example from both directionalities.

English-Arabic Directionality (Example Abstract 1)

The abstract under scrutiny delves into the thematic realm of Algerian Arabic poems known as "Buqalah," which encapsulate connotative insights into the resistance of Algerian women during the

French colonization era. In the comparative analysis of translation outputs across three different scenarios, a congruence of merely 22% was observed, with a notably higher alignment of 40% between the two LLMs (refer to Appendix B for an in-depth analysis).

A noteworthy observation is that Google Translate (GT) exhibited limitations in its translational capacity, particularly in the context of geographic nomenclature. Titles corresponding to city names such as *Buqalah, Blida, Cherchell, Tlemcen, Constantine,* and *Algiers* were left untranslated, retaining their original form in the source language (SL). In stark contrast, both ChatGPT 3.5 and 4 adeptly identified and accurately translated these names as Algerian city titles. Furthermore, GT demonstrated a deficiency in converting the grammatical functions of certain terms. For instance, the word "morality" (a noun in English) was erroneously rendered into an Arabic adjectival form شفوية by GT, and similarly mistranslated as شفهية by ChatGPT 3.0. ChatGPT 4, however, adeptly translated it into the noun شفاهية, which aligns more coherently with its usage in the Arabic academic genre.

Literal translations in GT were also evident, as seen in the translation of "oral literature" to الأدب الفموي, an idiosyncratic rendition, whereas the LLM tools offered more contextually accurate translations. Additionally, a masculine bias was observed in the outputs of GT and ChatGPT 3.5 in instances requiring a feminine pronoun, specifically in the phrase "يرتبط الطقوس." This gender-specific inaccuracy was notably absent in the translations provided by ChatGPT 4.

The translation of the term "divinatory" varied across the platforms, being rendered as إلهية, تنبؤ, and تنجيم in GT, ChatGPT 3.5, and ChatGPT 4, respectively. Notably, ChatGPT 4's translation encapsulates the contextual essence of the term more effectively, highlighting its advanced comprehension and translational capabilities.

Arabic-English Directionality (Example Abstract 6)

The abstract in question examines a Syrian novel, which explores the theme of alienation among the Syrian populace. A significant distinction was observed in the treatment of the novel's title by different translation tools. GT and ChatGPT 3.5 resorted to a mere transliteration of the title, whereas ChatGPT 4 demonstrated a superior capability by accurately translating the title into English.

In the realm of lexical translations, GT's approach occasionally deviated from conveying the intended meanings. For instance, the Arabic word “تحدثت” was translated by GT as “spoke”, which diverges from the contextually appropriate meaning. In contrast, ChatGPT 3.5 and 4 adeptly translated it as “address” and “discuss”, respectively, thus preserving the intended nuance. Similarly, the literal translation of “تأتي” as “come” by GT was identified as inadequate, a shortcoming that was effectively rectified in the translations provided by the LLM tools.

A notable discrepancy was also observed in the translation of the phrase “السرد الروائي.” GT rendered it as “narrative narration”, an awkward and redundant expression due to the repetition of similar words with only a slight grammatical variation. ChatGPT 3.5 and 4, however, proficiently translated it to “narrative structure”, thereby eliminating the redundancy and enhancing coherence. An additional aspect where GT’s limitations were apparent pertains to sentence structuring. GT tended to adhere to the syntactic rules of the source language (SL), often opting for comma separation over full stops, resulting in lengthy, run-on sentences. This sharply contrasted with the translations generated by the LLM tools, which respected the syntactic norms of the target language (TL), resulting in more segmented and coherent sentence structures. Specifically, ChatGPT 3.5 produced an abstract comprising five sentences, while ChatGPT 4 generated an abstract of six sentences.

The analysis of identical word matches across the three translation tools yielded a matching rate of 28%, while a substantially higher rate of 74% was observed between the AI tools. This disparity underscores the enhanced accuracy and consistency of AI-based translation tools compared to conventional tools like GT. Further details and analyses can be found in Appendix C.

Discussion

The primary objective of the present study was to evaluate the proficiency of conventional translation (CT) tools and advanced artificial intelligence (AI) translation tools in addressing the inherent challenges of translating academic abstracts from literary studies. The study aimed to ascertain the extent to which these tools could accurately identify and overcome the unique difficulties posed by such texts and measure the closeness of their translation outputs to human translation standards. The results revealed a notable deficiency in GT, a representative of CT tools, particularly in its ability to comprehend contextual words and nuances within academic abstracts that encompass literary styles. This inadequacy was observed in translations involving both SL to TL and vice versa, leading to outputs that were often unreadable and lacked coherence. A significant shortcoming of GT was its inability to accurately identify and translate many words in the TL, often resorting to mere transliteration. This limitation was especially pronounced in cases involving the titles of cities or locations, where GT failed to capture the intended meaning.

Furthermore, the study's findings partially resonate with the observations made by Toral and Way (2015). Their research highlighted that MT tools generally exhibit poor performance in translating literary texts, attributed primarily to the complex nature of literary genres and the inherent differences between language systems. This study corroborates their findings, particularly in the context of GT struggling with translations from Arabic to English. GT's translations were often constrained by the rules and stylistic elements of the SL, resulting in excessively long sentences or, in some cases, entire abstracts being condensed into a single sentence. This issue exemplifies GT's limitations in adapting to the structural and stylistic differences between the SL and TL. In contrast, LLM tools demonstrated superior performance in translating outputs, producing nearly coherent sentences while accurately preserving the intended meanings of the SL and ensuring readability in the TL. This can be attributed to the inherent capabilities of LLMs to comprehend the contextual nuances embedded within the SL texts and effectively convey the meaning into the target texts. The exceptional performance exhibited by LLM models can be attributed to extensive training aimed at enhancing translation outputs and overcoming the challenges encountered by LLMs. These concerted efforts have contributed significantly to refining the translation capabilities and improving the overall quality of LLM-based translations (Lyu et al., 2023).

Our findings indicate that all of the LLM tools showcased a significant advantage over the translation outputs generated by GT for both Arabic and English academic abstracts. Notably, inaccuracies were observed in the translation outputs of GT and ChatGPT 3.5 pertaining to the feminine pronoun. Additionally, ChatGPT 3.5 exhibited shortcomings in recognizing the neutral pronoun "they", which can be attributed to the limited availability of training data for low-resourced languages like Arabic. These limitations resulted in certain inaccuracies in the translation outputs and occasional instances of hallucination. Arabic, similar to Bangali, has been identified as a low-resourced language in previous studies, which may explain the difficulties faced by ChatGPT 3.5 in accurately identifying pronouns. However, our research highlights that ChatGPT 4 demonstrates remarkable competence in recognizing pronouns and subject antecedents for both Arabic and English translations, overcoming the aforementioned challenges presented by its predecessor (ChatGPT 3.5), as well as GT. The findings of our study align with those of Sanz-Valdivieso and López-Arroyo (2023) providing further support for their observations. They reported that ChatGPT 3.5 generated fewer significant errors compared to GT when translating vocabulary related to olive oil and wine. These results reinforce the notion that ChatGPT 3.5 exhibits improved accuracy and proficiency in handling domain-specific terminology within this particular context. This is also in line with Khoshafa's (2023) it becomes evident that while ChatGPT 3.5 has shown improvements in certain areas, it is not without its limitations. The study by Khoshafa points out that, despite advancements, ChatGPT 3.5 still encounters difficulties in fully comprehending complex texts and accurately identifying terminologies that have cross-cultural implications. This suggests that while ChatGPT3.5 has made strides in domain-specific translation, challenges remain in its ability to consistently understand and translate texts that involve intricate conceptual nuances or that bridge diverse cultural contexts.

In the comparative assessment of LLMs, specifically ChatGPT 3.5 and ChatGPT 4, a discernible superiority in performance was noted for ChatGPT 4 in the context of translating Arabic texts into English. This heightened efficacy can be attributed to a multitude of factors, central to which is the ability of ChatGPT 4 to preserve the intended meaning, maintain readability, and adhere to the academic style of the TL. This observation is congruent with the research conducted by Son and Kim (2023), which highlighted that translations from non-English languages to English by ChatGPT 4 were more proficient compared to other MT tools such as GT and Microsoft Translator. Further supporting this conclusion are the findings of Lyu et al. (2023), who noted that translations performed by ChatGPT 3.5 for low-resourced languages occasionally resulted in 'hallucinations' or significant inaccuracies.

This phenomenon underscores the challenges inherent in translating languages with limited resources and representation in global linguistic databases. In contrast, ChatGPT 4's enhanced performance can be largely ascribed to the continuous evolution and improvements in AI technology aimed at simulating human-like translation capabilities. As elucidated by Ray (2023), key advancements contributing to this progression include a more nuanced understanding of context, the reduction of biases, and refined fine-tuning capabilities. The non-identical nature of the outputs from both ChatGPT 3.5 and ChatGPT 4, especially in terms of matching rates (as detailed in the supplementary materials), further illustrates the evolutionary leap from the former to the latter. The high evaluation scores assigned by raters to ChatGPT 4's translations attest to its remarkable ability to closely mirror human translation. This is particularly evident in its proficiency at capturing the contextual meanings of the SL text and effectively navigating the complexities of literary translation.

Conclusion and Limitations

The study presented herein focuses on assessing the accuracy and effectiveness of three translation tools – GT, which utilizes NTM, and two LLMs, ChatGPT 3.5 and 4 – in translating academic abstracts within the literary genre. Our findings reveal a significant disparity in performance among these tools. GT's outputs were generally inaccurate, often failing to grasp contextual phrases, resulting in unreadable texts, and heavily influenced by the style of the SL. Conversely, the LLMs, represented by ChatGPT 3.5 and 4, demonstrated remarkable aptitude in processing literary texts, adhering closely to the academic writing conventions of the target language (TL), and significantly outperforming GT.

Several factors underpin ChatGPT 4's superior performance. These include its extensive training dataset and advanced algorithms, which are fine-tuned to understand diverse genres and handle low-resourced languages effectively. Additionally, ChatGPT 4 integrates updates that refine its ability to mitigate biases and errors, particularly in underrepresented languages like Arabic. The implications of our findings could be particularly valuable for translator training, offering insights into how the efficiency of various translation modes can be leveraged to augment the human translation process. It is important, however, to recognize that despite the advancements of LLM tools, human intervention remains crucial to refine the outputs and ensure they accurately convey the intended meaning of the SL.

The study, however, is not devoid of limitations. Firstly, the analysis was based on a sample of only 10 abstracts for each language directionality. This limitation was due to the challenges in assigning voluntary raters to evaluate longer texts across three different tools. Future research might benefit from a larger sample size to yield more reliable and generalizable findings. Secondly, the study's focus on literary academic abstracts may not fully encapsulate the capabilities of these tools in other genres, which also warrant examination. Lastly, the study's scope was restricted to a select few translation tools. For a more comprehensive understanding of the capabilities of MT and LLMs, future studies should consider incorporating a broader array of translation tools to evaluate their efficacy in producing high-quality translation outputs.

ADHERENCE TO ETHICAL STANDARDS

Ethic declarations. The study does not involve human participants or animals. Therefore, no IRB application was required.

Data Availability. The data supporting the findings of this study are available within the manuscript. Raw datasets generated during the current study are available from the corresponding author, Dr. Mohammed Ali Mohsen, upon reasonable request.

Funding. This research received grant no. (15/2023) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

Conflicts of Interest. The authors declares no conflict of interest.

Consent for Publication. All authors commented on previous versions of the manuscript. All authors have read and approved the final version of the manuscript.

References

Abashev, A., Grigoryev, R., Grigorian, K., & Boyko, V. (2017). Programming Tools for Messenger-Based Chatbot System Organization: Implication for Outpatient and Translational Medicines. *BioNanoScience, 7*(2), 403–407. https://doi.org/10.1007/s12668-016-0376-9

Baidoo-Anu, D., & Ansah, L.O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Bulchand-Gidumal, J., William Secin, E., O'Connor, P., & Buhalis, D. (2023). Artificial intelligence's impact on hospitality and tourism marketing: exploring key themes and addressing challenges. Current Issues in Tourism. https://doi.org/10.1080/13683500.2023.2229480

Genzel, D., Uszkoreit, J., & Och, F. (2010). "Poetic" statistical machine translation: rhyme and meter. In Hang Li, & Lluís Màrquez (Eds.), *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (Massachusetts, USA, 9–11 October 2010) (pp. 158–166). Cambridge, MA. Association for Computational Linguistics.

Ghosh, S., & Caliskan, A. (2023). ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *arXiv preprint arXiv:2305.10510*. https://doi.org/10.1145/3600211.3604672

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., & Awadalla, H.H. (2023). How good are gpt models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210.*

Hossain, K.A. (2023). Analysis of Present and Future Use of Artificial Intelligence (AI) in Line of Fourth Industrial Revolution (4IR). *Scientific Research Journal, XI*(VIII), 1–50. http://dx.doi.org/10.31364/SCIRJ/v11.i8.2023.P0823954

Hurtado Albir, A., & Taylor, P. (2015). The acquisition of translation competence. Competences, tasks, and assessment in translator training. *Meta, 60*(2), 256–280. https://doi.org/10.7202/1032857ar

Hutchins, J. (2005). Example-based machine translation: a review and commentary. *Machine Translation, 19*(3), 197–211. https://doi.org/10.1007/s10590-006-9003-9

Islam, M.A., Anik, M.S.H., & Islam, A.B.M.A.A. (2021). Towards achieving a delicate blending between rule-based translator and neural machine translator. Neural *Computing and Applications, 33*(18), 12141–12167. https://doi.org/10.1007/s00521-021-05895-x

Jiao, W., Wang, W., Huang, J.-t., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745.*

Jones, F.R. (2019). *Literary translation*. Routledge encyclopedia of translation studies. https://doi.org/10.4324/9781315678627-63

Khoshafah, F. (2023). ChatGPT for Arabic-English translation: Evaluating the accuracy, 13 April 2023, PREPRINT (Version 1) available at Research Square. https://doi.org/10.21203/rs.3.rs-2814154/v1

Lyu, C., Xu, J., & Wang, L. (2023). New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181.*

Malik, T., Dwivedi, Y., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., & Raghavan, V. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy.

International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Mohsen, M.A., Althebi, S., & Albahooth, M. (2023). A scientometric study of three decades of machine translation research: Trending issues, hotspot research, and cocitation analysis. *Cogent Arts & Humanities, 10*(1). https://doi.org/10.1080/23311983.2023.2242620

Ponzio, A. (2007). Translation and the literary text. *TTR, 20*(2), 89–119. https://doi.org/10.7202/018823ar

Quah, C.K. (2006). *Translation and technology*. Springer. https://doi.org/10.1057/9780230287105

Ray, P.P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3,* 121–154. https://doi.org/https://doi.org/10.1016/j.iotcps.2023.04.003

Sanz-Valdivieso, L., & López-Arroyo, B. (2023). Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation? *Proceedings of the International Conference HiT-IT 2023 (Naples, Italy, 7–9 July 2023)* (pp. 97–107). https://doi.org/10.26615/issn.2683-0078.2023_008

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A.V.M., & Mokry, J. (2017). Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357.*

Son, J., & Kim, B. (2023). Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. *Information, 14*(10), 574. https://doi.org/10.3390/info14100574

Suta, P., Lan, X., Wu, B., Mongkolnam, P., & Chan, J.H. (2020). An overview of machine learning in chatbots. *International Journal of Mechanical Engineering and Robotics Research, 9*(4), 502–510. https://doi.org/10.18178/ijmerr.9.4.502-510

Toral, A., & Way, A. (2015). Translating literary text between related languages using SMT. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* pages *(Denver, Colorado, USA, June 4, 2015)* (pp. 123–132). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0714

Valdeón, R.A. (2023). Automated translation and pragmatic force: A discussion from the perspective of intercultural pragmatics. *Babel, 69*(4), 447–464. https://doi.org/https://doi.org/10.1075/babel.00328.val

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A.N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., & Parmar, N. (2018). Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416.*

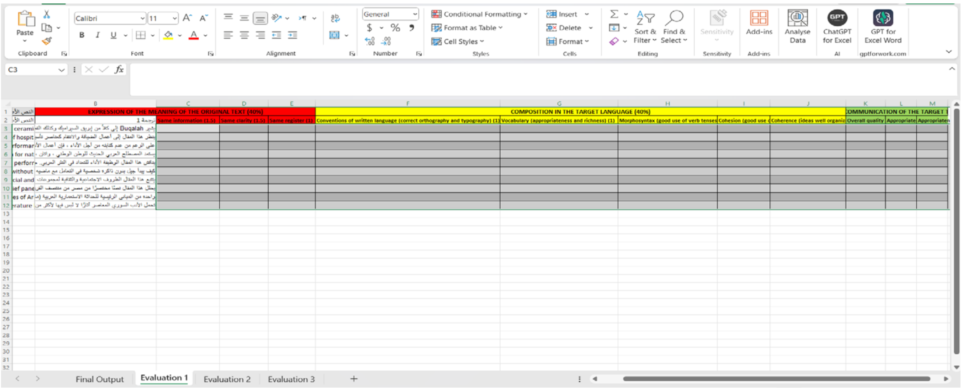
Voigt, R., & Jurafsky, D. (2012). Towards a literary machine translation: The role of referential cohesion. David Elson, Anna Kazantseva, Rada Mihalcea, Stan Szpakowicz (Eds.), *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature (Montreal, Canada, June 8, 2012)* (pp. 18–25). Association for Computational Linguistics.

Wu, E.H.-K., Lin, C.-H., Ou, Y.-Y., Liu, C.-Z., Wang, W.-K., & Chao, C.-Y. (2020). Advantages and constraints of a hybrid model K-12 E-Learning assistant chatbot. *Ieee Access, 8*, 77788-77801. https://doi.org/10.1109/ACCESS.2020.2988252

Wu, L., Tian, F., Qin, T., Lai, J., & Liu, T.-Y. (2018). A study of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1808.08866.*

Appendix A

Screenshot of the Evaluation Sheets



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	النص الأ	EXPRESSION OF THE MEANING OF THE ORIGINAL TEXT (40%)				COMPOSITION IN THE TARGET LANGUAGE (40%)					OMMUNICATION OF THE TARGET T		
2	النص الأ	ترجمة 1	Same information (1.5)	Same clarity (1.5)	Same register (1)	Conventions of written language (correct orthography and typography) (1)	Vocabulary (appropriateness and richness) (1)	Morphosyntax (good use of verb tenses	Cohesion (good use o	Coherence (ideas well organiz	Overall quality	Appropriate	Appropriaten
3	cerami	يشير Buqalah إلى كلاً من إبريق السيراميك وكذلك الت											
4	f hospit	ينظر هذا المقال إلى أعمال الضيافة والانتقام كعناصر تأسي											
5	rformar	على الرغم من عدم كتابته من أجل الأداء ، فإن أعمال الأ											
6	for nat	يستمد المصطلح العربي الحديث للوطن الوطني ، واثان ،											
7	perforr	يناقش هذا المقال الوظيفة الأداء للتعداد في النثر العربي.											
8	without	كيف يبدأ جيل بدون ذاكره شخصية في التعامل مع ماضيه											
9	cial and	يتتبع هذا المقال الظروف الاجتماعية والثقافية لمجموعات											
10	ief pane	يحلل هذا المقال نصًا مختصرًا من مصر من منتصف القر											
11	es of Ar	واحده من المباني الرئيسية للحداثة الاستعمارية العربية (م											
12	erature	تحمل الأدب السوري المعاصر آثارًا لا لبس فيها لأكثر من											

Appendix B

Translation Comparison Table with Match Rates (Updated Data)

https://nejranuniversity-my.sharepoint.com/:u:/g/personal/mamohsen_nu_edu_sa/Eb6Cig9mnUlHo2xLjPACBqsB79vthdWvtTV3eGRlj5KaXw?e=vN752m

Appendix C

Translation Comparison Table with Match Rates

https://nejranuniversity-my.sharepoint.com/:u:/g/personal/mamohsen_nu_edu_sa/ERDoHtPwcYlHmEcuhCBFk_8BpkN89goT8PLbb5IDURvN8Q?e=6IV1uf

АНОТАЦІЯ

Мета. *Поява в листопаді 2022 року генеративної моделі штучного інтелекту (ШІ) Large Language Model (LLM) справила глибокий вплив на різні сфери, зокрема й на перекладознавство. Це спонукало нас провести ретельну оцінку ефективності й точності машинного перекладу, представленого Google Translate (GT), у порівнянні з великими мовними моделями (LLM), зокрема ChatGPT 3.5 і 4, при двосторонньому перекладі академічних рефератів з англійської та арабської мов.*

Методи. *Застосовуючи змішаний підхід, у цьому дослідженні використано корпус, що складається з 20 рефератів, взятих з рецензованих журналів, індексованих у Clarivate Web of Science, зокрема, Journal of Arabic Literature та Al-Istihlal Journal. Анотації розділені порівну, щоб представити як англо-арабський, так і арабсько-англійський напрямки перекладу. Дизайн дослідження ґрунтується на комплексній шкалі оцінювання, адаптованій з Hurtado Albir and Taylor (2015), з акцентом на семантичну цілісність, синтаксичну зв'язність і технічну адекватність. Троє незалежних експертів оцінювали результати перекладу, отримані за моделями GT і LLM.*

Результати. *Результати кількісного та якісного аналізу показали, що інструменти LLM значно перевершують результати перекладу за допомогою MT як в арабському, так і в англійському напрямках. Крім того, ChatGPT 4 продемонстрував значну перевагу над ChatGPT 3.5 в арабсько-англійському перекладі, тоді як в англо-арабському напрямку перекладу статистично значущої різниці не спостерігалося. Результати якісного аналізу показали, що інструменти ШІ здатні розуміти контекстуальні нюанси, розпізнавати назви міст і адаптуватися до стилю цільової мови. І навпаки, ГТ демонстрував обмеження в роботі з конкретними контекстуальними аспектами і часто надавав дослівний переклад певних термінів.*

Ключові слова: *ChatGPT, машинний переклад, Google Translate, анотація статті.*