ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 16, No. 6, pp. 2025-2035, November 2025
DOI: https://doi.org/10.17507/jltr.1606.24

A Comparative Study of AI-Powered Tools for Arabic-English and English-Arabic Translation

Razan R. Khasawneh
Department of English Language and Translation, Amman Arab University, Jordan

Bilal I. Alsharif
Department of Applied Linguistics, Al-Zaytoonah University of Jordan, Jordan

Abstract—**This study is a quantitative analysis that investigates and compares the quality level of Computer Assisted Translation (CAT) tools, Neural Machine Translation (NMT) systems, and Large Language Models (LLMs) against each other and how well they can perform on translation tasks from English into Arabic and vice versa, in comparison to human translator. By utilizing a Bilingual Evaluation Understudy (BLEU) and comparing seven translation platforms, the study revealed that AI models outperform both CAT tools and NMT systems when translating in both directions. In addition, Gemini scores the highest BLEU score in both directions, surpassing all the other six platforms, while Google Translate scores are the lowest. The study also reveals that these platforms struggle more with English-to-Arabic translation due to the complexity of the Arabic language. Since CAT tools generally have the lowest scores, they assist the human translator rather than replacing them, providing a partially automated translation. Accordingly, AI models can be more helpful and preferable due to their high BLEU score. However, despite often producing accurate translations and conveying the core meaning, MT may still struggle to imitate the stylistic choices and conciseness that characterize a professional human translator at work.**

Index Terms—**Computer Assisted Translation (CAT), Neural Machine Translation (NMT), Large Language Models (LLMs), Bilingual Evaluation Understudy (BLEU), Machine Translation (MT)**

I. Introduction

Technological advancements have long helped shape and improve different aspects of today's world, from communication and healthcare to industry and education, including translation. To meet market demands, numerous efforts have been made to create tools and software that assist translators in becoming more efficient and productive. These efforts result from collaboration between translators, linguists, software developers, and AI researchers. Machine translation (MT) and Computer Assisted Translation (CAT) technologies are two significant technological advancements that have drastically altered communication (Doherty, 2016). Hutchins (1995) states that MT involves computerized systems that produce translations with or without human assistance, excluding CAT tools that aid translators by providing various features. It is the automation of the whole translation process. Beyond CAT tools, developers created other systems, including Neural Machine Translation (NMT) and AI-powered platforms that utilize Large Language Models (LLMs). NMT, such as Google Translate, seeks to create a single neural network that can be jointly adjusted to optimize translation efficiency (Bahdanau et al., 2014) using a neural network approach. Large volumes of text data are used to train LLM models, such as ChatGPT and DeepSeek. These systems can produce human-like text, accurately answer queries, and perform other language-related tasks (Kasneci et al., 2023).

As MT software and technologies continue to develop, consistent and thorough evaluation has become essential to determine and ensure that the MT software meets the quality standards of accuracy, fluency, and adequacy. According to Han (2022), evaluating statistical and NMT to identify the reliability and limitations of similar systems, including CAT tools and LLM models, is important. We are particularly interested in identifying the quality level of each tool and how well NMT, LLMs, and CAT tools can perform on translation tasks compared to human translators by utilizing a Bilingual Evaluation Understudy (BLEU). In addition, the study aims to identify which of these MT platforms surpasses the others in their ability to translate between Arabic and English. BLEU measures the degree to which machine-translated text matches one or more human reference translations (Kleinman, 2023). A high BLEU score indicates a high similarity between the generated computer text and the human reference translation. Accordingly, Google Translate and Microsoft Translate are used as NMT software, while ChatGPT-4, DeepSeek, and Gemini are used to represent LLMs. Additionally, Smartcat and Matecat are used to represent CAT tools. This study compares Arabic-English translation and English-Arabic translation. Al-Kabi et al. (2013, p. 66) state, "Due to its rich and complex morphological features, Arabic has always been challenging for machine translation. In addition, Arabic has different word forms and word orders, which make it possible to express any sentence in different forms".

II. Literature Review

A. Machine Translation

Machine translation (MT) is the automatic translation of a text or speech from one language into another using computer technology. It is a branch of applied linguistics and information technology that involves translating human languages (Abiola et al., 2015), which can occur with or without human assistance. Correspondingly, it may include post-editing, which is the human revision of the machine-translated output after the MT system has finished its task, or pre-processing, which is the preparation of the text for the MT system without affecting its linguistic analysis capabilities (Alsharif et al., 2025). One of the primary goals of machine translation is to develop a system that consistently produces high-quality translations between human languages (Adeoye, 2012). However, the output of MT is usually revised (post-edited), which is similar to the output of most human translators, who typically revise it with a second translator (Hutchins, 1995). MT encompasses numerous approaches, including a direct approach that attempts to generate a translation straight from the source language (SL) to the target language (TL) without the use of an intermediate representation (Poibeau, 2017), a rule-based approach using morphological, syntactic, and semantic rules to analyze the SL text and synthesize the TL content (Abiol et al., 2015); corpus based approach which includes statistical machine translation (SMT) and example-based machine translations (EBMT); and the knowledge-based approach that makes use of a broad range of pragmatic and semantic expertise and can reason about concepts (Sharma, 2011). In addition, hybrid MT combines multiple approaches, offering improved translation quality and functionality compared to traditional methods, but is computationally more complex (Dhariya et al., 2017). This approach includes a word-based model, a phrase-based model, a syntax-based model, and a forest-based model.

There are several advantages to the translation generated using MT systems. It offers speed and cost-effective translation compared to manual one. Moreover, it enhances consistency, particularly in terminology, and is suitable for translating online content (Abdulsalami & Akinsanya, 2017). In contrast, it is considered to be a complex task due to the complex nature of languages, including word ambiguity, differences in sentence structure, and the need for contextual knowledge (Abiola et al., 2015). In addition, some aspects of the language are culturally bound and not easy to understand, which might cause a problem in the translation process (Khasawneh et al., 2025).

B. Machine Translation Quality Assessment

Evaluating MT systems is crucial for determining the quality and efficiency of the translated output. In industry, evaluation focuses on the final product or customer, whereas in research, evaluation aims to demonstrate significant improvements compared to previous research or alternative translation methods (Castilho et al., 2018). There are two methods for judging and evaluating MT outputs: human or manual evaluation and automatic evaluation. This procedure is crucial to guarantee accuracy, fluency, and sufficiency whether the translation is produced by a machine or by a human (Alsharif & Khasawneh, 2025).

Experts in linguistics and translation evaluate human evaluation in terms of adequacy and fluency (Maučec & Donaj, 2019). Adequacy evaluates semantic qualities, whereas fluency evaluates syntactic qualities. Human evaluation methods include error identification, annotation, and classification (Rivera-Trigueros, 2022). Correspondingly, evaluators require proficiency in both SL and TL. This method has several drawbacks: it is time-consuming, has a substantial cost, is inflexible, and is not repeatable. Graham (2015) notes that human evaluation of translation quality offers the most insightful assessment of systems, yet human assessors are notoriously unreliable, which makes quality estimation difficult. Since human or manual evaluation has some drawbacks, automatic assessment metrics have been frequently employed for machine translation.

In the automatic evaluation method, the output of the MT system is compared with one or more reference translations (Castilho et al., 2018; Han, 2016), i.e., the quality of MT should be close to human quality. Having multiple human reference translations for every machine-translated sentence being assessed is crucial because even human translation might vary greatly (Maučec & Donaj, 2019). However, some metrics, such as Quality Estimation (QE) and intrinsic evaluation metrics, do not use reference translation. MT evaluation uses basic metrics based on word overlap, sequence similarity, word order, and edit distance. In contrast, advanced metrics incorporate linguistic features such as syntax (POS, sentence structure) and semantics (entailment, paraphrase, synonyms, named entities, semantic roles), as well as language models (Graham, 2015). Several metrics are used to evaluate MT output, including NIST, METEOR, TER, and Bilingual Evaluation Understudy (BLEU), which are the most widely used.

BLEU primarily uses n-gram precision, which refers to the n-grams in the MT that can also be found in the reference translation. It evaluates multiple n-gram lengths (up to four words) and combines the results using a geometric average; higher-order n-grams directly assess the grammatical accuracy of the translation (Maučec & Donaj, 2019). In addition, BLEU applies a brevity penalty to penalize translations that are too short compared to the reference. The BLEU metric ranges from 0, indicating "almost useless translation", to 1, indicating a "perfect match" (Papineni et al., 2002). Unless a translation is identical to a reference translation, very few translations receive a score of 1. This is because there is a wide range of possible accurate translations; even a human translator will not always receive a perfect score. BLEU has some drawbacks; n-gram matching requires exact word matches, it struggles to differentiate between low-quality translations, as its geometric mean calculation of modified precision scores results in a zero score if any individual n-gram precision is zero, it equally rewards all matches of the same length, disregarding the semantic importance or content value of the

matched text, and finally, the reliability of BLEU scores is heavily dependent on the number of human reference translations used, ideally requiring comprehensive coverage of all possible valid translations for a given source sentence to achieve optimal accuracy (Chatterjee et al., 2007).

C. CAT Tools

Computer-assisted translation (CAT) tools are software that help translators convert text from one language into another more quickly and efficiently. They primarily store the translated text units in the translation memory (TM) database, which can be used later in other texts or within the exact text (Mukhtar, 2025) to maintain consistency and improve quality. The translation memory principle in TMs is applied to segments, not individual phrases. However, TM has a bad reputation for "sentence salad" (Bédard, 2000), “peep-hole translation” (Heyn, 1998), and “blind faith” (Bowker, 2005). According to Manojlovic et al. (2020), the first occurs when overused or mismatched segments disrupt the flow and coherence of a translation. The second is due to ignoring cross-sentence cohesion, such as anaphora, by focusing only on sentence-level matches. The last refers to the tendency to use TM matches non-critically. Therefore, translators need to review the text and check any suggestions made by the system. The CAT tool's drawbacks also include a learning curve, high cost, over-reliance on TM, and performance issues (Mukhtar, 2025).

Nevertheless, CAT tools are still commonly used among professional translators due to their quality assurance, consistency, increased productivity, multi-format friendly, and concordance, which are considered among CAT tools' advantages. Smartcat, an all-in-one cloud-based platform, and Matecat, an open-source online CAT tool, are leading CAT tools. Both can be used for free and offer advanced features, including translation memory, integrated MT, client services, and training. However, Smartcat's offline functionality is limited; it does not offer specific editing choices, and using other functions may incur additional costs, while Matecat may be slow at the final stage of the project (Karpina, 2024). Moreover, Matecat needs a permanent internet connection, but it might lack some advanced features available in paid tools, face consistency issues, and have a file size limit.

D. Neural Machine Translation

Neural machine translation (NMT) is the process of converting text from one language to another using an artificial neural network (ANN). NMT learns the model jointly to maximize translation performance through a two-step recurrent neural network (RNN) consisting of an encoder and a decoder, which are AI translation platforms (Bahdanau et al., 2014; Wołk & Marasek, 2015). According to Cho et al. (2014), a variable-length input sentence is converted into a fixed-length vector representation by the encoder, and the decoder uses this representation to produce an accurate variable-length target translation. In other words, the NMT model receives the ST, reads it, and, based on its own understanding of the sentence, generates the target sentence word by word (Wanga et al., 2022).

NMT excels at learning the direct mapping between input and output texts in an end-to-end manner, avoiding the complex design of traditional phrase-based systems (Wu et al., 2016). However, NMT faces accuracy challenges, particularly with large datasets. NMT drawbacks include slower training and inference, difficulty handling rare words, potential translation omissions, or failure to translate all words. Arthur et al. (2016) state that NMT frequently makes mistakes when translating low-frequency content terms, which are essential for comprehending the sentence's meaning. Google Translate and Microsoft Translate utilize NMT technology, which continues to evolve, is user-friendly, and offers fast and offline translation across a wide range of languages through their mobile applications. However, Google Translate lacks contextual understanding, ignores regional dialects, struggles with nuanced phrases, does not always follow the correct syntax for every language, over-reliance on Machine Learning, and over-simplification of Complex Texts (Raymond, 2024), which also applies to Microsoft Translate.

E. Large Language Models

Large Language models (LLMs) are a form of artificial intelligence algorithm trained on a vast amount of data to generate and understand human languages or texts with minimal human intervention. Based on natural language instructions, recent developments in LLMs have demonstrated impressive zero-shot capabilities across text-generation tasks (Chung et al., 2024). In addition to having a substantially larger model size, LLMs demonstrate superior language creation and comprehension skills and—more significantly—emergent skills absent from smaller-scale language models (Minaee, 2025).

Out of all the tasks LLMs can perform, such as writing documents, creating executable code, and answering questions, often with human-like capabilities (Schulman et al., 2022), translation has become one of the most notable areas where LLMs have demonstrated remarkable ability and proficiency (Agrawal et al., 2022; Moslem et al., 2023; Peng et al., 2023). The use of LLMs for automatic code translation has excellent potential, as LLMs with more parameters typically exhibit stronger translation capabilities; however, LLMs' accuracy still suffers from accuracy issues (Yang et al., 2024). LLMs must be able to comprehend both code syntax and translation semantics concurrently, and a translation is deemed successful if it passes run-time checks and pre-existing tests on the translated code (Pan et al., 2024). AI can transform various aspects of our lives, including industries, education, and healthcare. In the realm of translation, AI is a game-changer. It has reshaped translation practices, training, and the industry as a whole. Some well-known LLMs include ChatGPT (OpenAI GPT), DeepSeek AI (Open-Source), and Gemini from Google DeepMind.

ChatGPT is a generative artificial intelligence chatbot developed by OpenAI, which follows instructions and provides

more conversational, comprehensive responses to questions. It can serve as a writing assistant, providing advice, job interview preparation, and language translation. As a result, ChatGPT is regarded as a turning point for AI (Mollick, 2022) and has even raised a "code red" alarm for NMT (Khan, 2022), such as Google. It has an advanced Generative Pre-trained Transformer (GPT) architecture that understands and produces human-like text with unprecedented fluency (Bender et al., 2021). While ChatGPT has numerous advantages, it also has several limitations, including accuracy issues, dependence on the quality of input, a lack of human touch, and privacy concerns (Mukhtar, 2025).

DeepSeek is a Chinese open-source Vision-Language (VL) Model built upon the DeepSeek language model series. It provides competitive performance and significantly reduces the cost of inference and training (Bi et al., 2024). A key characteristic of DeepSeek's AI model is its efficiency and affordability. It excels exceptionally well on assignments that require technical problem-solving, precise grammar, and structured writing (Joshi, 2025). It is used for many of the same tasks as GhatGPT, including translation. However, its effectiveness compared to that of its rivals remains an open question.

Gemini is the model developed by Google. It includes different versions, including Ultra, Pro, and Nano, each designed with distinct capabilities and intended uses. Ultra for maximum power on complex tasks, Pro for a versatile balance across many tasks, and Nano for efficient on-device AI. Their architecture is based on transformer decoders, which have been enhanced through model optimization and architectural improvements to enable reliable training at scale (Algobaei et al., 2025). Gemini also performs several tasks, including text summarizing, reasoning, and multilingual tasks for translation. Correspondingly, Team et al. (2023) note that Gemini showed remarkable performance when translating from English to other languages and surpasses other LLM-based translation methods. It can create fluent, human-like, genuine translations; however, human intervention is still essential to ensure accuracy, particularly in high tasks and professional settings (Ferrag & Bentounsi, 2024). According to Al-Salman and Haider (2024), Gemini can make translation errors, including transliterating terms rather than providing their correct translated equivalent.

III. Methodology

This study adopts a quantitative research methodology to compare and evaluate the translation precision for CAT tools, NMT systems, and LLMs. Correspondingly, Smartcat and Matecat are used as CAT tools, while Google Translate and Microsoft Translate are used as NMT software. Additionally, ChatGPT-4, DeepSeek, and Gemini are utilized as LLM platforms. These platforms are chosen for their advanced AI capabilities and popularity. The study sample is collected from Tatoeba (2025), an open-source website with an extensive database of sentences and their translations into other languages. It has been chosen because it provides a diverse range of sentence structures, vocabulary, and topics. To obtain a valid and meaningful measurement, 1000 sentences translated from Arabic into English and another set of 1000 sentences translated from English into Arabic are randomly selected to conduct the study. The BLEU metric has been utilized to evaluate the similarities between human-generated translations, reference translations, and machine translations. It has been employed because it is the most popular metric, as Papineni et al. (2002) is one of the most cited papers in NLP. Both reference translation and MT inputs undergo consistent preprocessing, including tokenization and normalization. Tokenization involves breaking sentences into individual words, and normalization entails converting all text to lowercase. Further, punctuation and special characters are removed to ensure a more accurate and fair evaluation. Then, Python, a high-level, general-purpose programming language, is used to calculate BLEU. The following is the BLEU metric equation for calculating the BLEU score:

$$\text{BLEU} = \text{BP} \times \exp\left(\frac{1}{n} \sum_{i=1}^{n} \log p_i\right)$$

In addition, researchers included a breakdown of the BLEU score calculation by n-gram (1-gram, 2-grams, 4-grams, and 4-grams). This provides more granular insights into fluency and adequacy of the translation at different levels of word sequence matching. Boxplots, generated using R for statistical and data visualization, are also used to visually compare the translation quality between the translation platforms of the same type based on the BLEU score. The X-axis represents the two machine platforms, while the Y-axis displays the BLEU score. The line within the box indicates the median BLEU score. Follow-up examples are provided to highlight the subtle differences between the seven platforms. The length ratio, which complements BLEU, which applies a brevity penalty to penalize translations that are too short compared to the reference, is also provided to compare the length of the reference translation to the length of the MT output, as it measures fluency and competence, while BLEU measures accuracy. The length ratio can be interpreted as follows:

- Length ratio < 1: MT output is shorter than the reference translation.
- Length ratio = 1: MT output has the same length as the reference transition.
- Length ratio > 1: MT output is longer than the reference translation.

IV. Results and Discussion

This section presents the results of machine translation evaluation based on the BLEU metric. Seven translation platforms were selected to test and compare their translation performance in translating the selected data from Arabic into English and vice versa. Additionally, N-gram BLEU scores, ranging from 1 to 4 grams, are also presented. Higher BLEU scores generally indicate better translation quality, i.e., MT output is closer to the translation produced by a human translator.

Table 1 below shows BLEU scores for the seven machine translation platforms when translating from English into

Arabic. After evaluating 1,000 sentences and comparing MT engines, Gemini and ChatGPT-4 top the BLEU score across all platforms, with overall BLEU scores of 41.30 and 40.94, respectively. This is followed by DeepSeek, with a score of 39.90, closely aligned with Smartcat (39.58), Google Translate (39.48), and Microsoft Bing (39.41). Matecat has the lowest score, with 38.53.

TABLE 1
BLEU SCORE OF ENGLISH-ARABIC DATA

	CAT Tools		NMT		AI		
Type	Smartcat	Matecat	Google Translate	Microsoft Bing	Gemini	ChatGPT-4	DeepSeek
1-Gram	74.01	73.65	73.38	73.37	75.68	75.87	75.02
2-Gram	49.75	48.69	49.69	49.67	52.64	52.24	51.41
3-gram	32.75	31.45	32.41	32.33	35.06	34.71	33.60
4-gram	22.12	21.31	21.78	21.72	24.27	23.84	22.24
BLEU	39.58	38.53	39.48	39.41	41.30	40.95	39.90

Figure 1 compares the performance of CAT tools in terms of the overall BLEU score. Smartcat performs slightly better than Matecat, with scores of 39.58 and 38.53, respectively. Matecat's n-gram scores are also lower than Smartcat's.

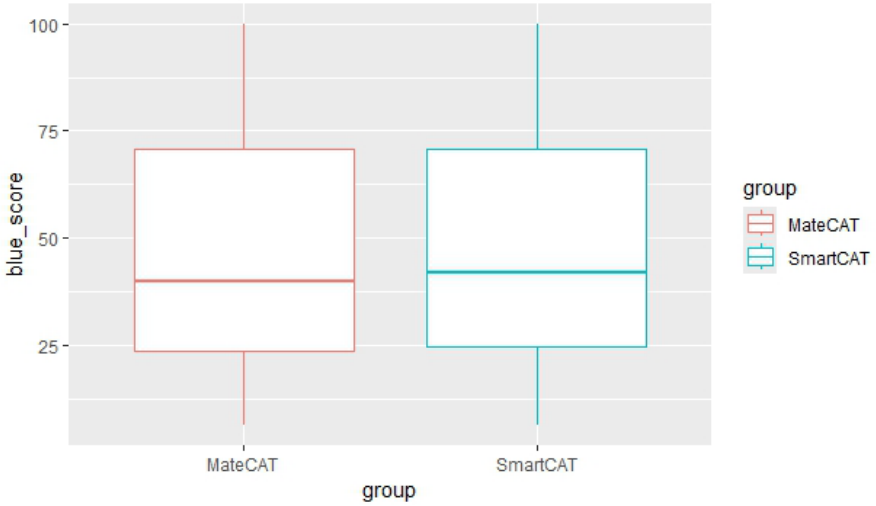

Figure 1. Boxplot of BLEU Score of Matecat and Smartcat (English-Arabic)

Figure 2 presents the BLEU scores of NMT. Accordingly, Google Translate's overall BLEU score is 39.48. Meanwhile, Microsoft Bing scores similarly to Google Translate, with an overall score of 39.41. Their n-gram performance is also very close.

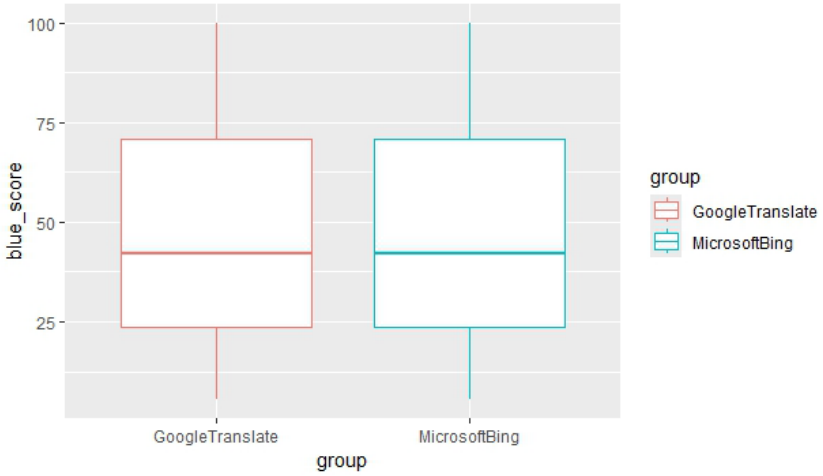

Figure 2. Boxplot of BLEU Score of Google Translate and Microsoft Bing (English-Arabic)

Regarding AI systems known for their advanced capabilities, Figure 3 indicates that Gemini has scored an overall BLEU score of 41.30. Gemini also performs strongly in all n-gram levels, particularly in 1-gram, with a score of 75.68. GhatGPT and DeepSeek have scored an overall BLEU score of 40.95 and 39.90, respectively. ChatGPT-4 n-gram scores are also competitive, with 1-gram precision (75.68) to Gemini. However, DeepSeek n-gram scores are consistently lower than those of Gemini and ChatGPT-4.

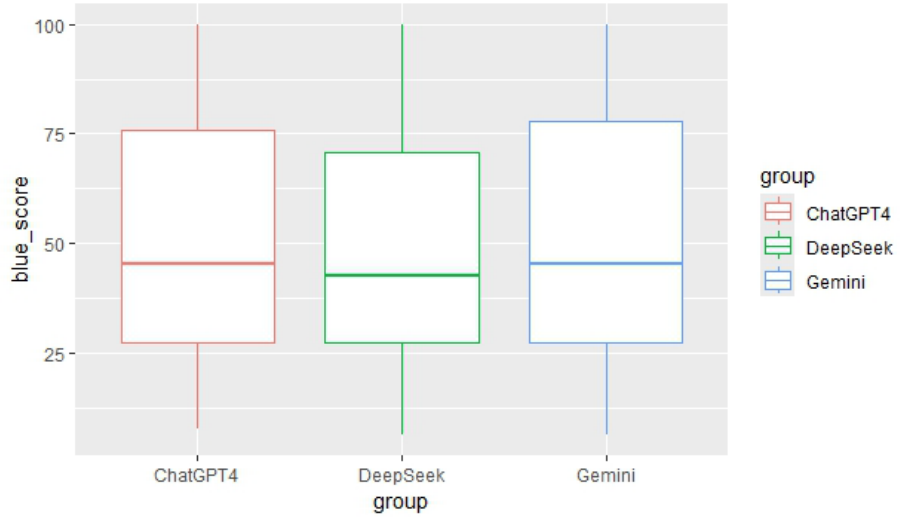

Figure 3. Boxplot of BLEU Score of ChatGPT-4, DeepSeek, and Gemini (English-Arabic)

Table 2 presents BLEU scores for the seven machine translation platforms under study when translating from English into Arabic. It is worth noting that the performance of all platforms for the English-to-Arabic dataset has improved. Gemini leads all platforms with the highest overall BLEU score of 59.73, followed closely by DeepSeek and ChatGPT-4, with BLEU scores of 59.12 and 57.95, respectively. Next comes Smartcat, with a score of 55.69, which is higher than Matecat's score of 54.68. The lower-performing platforms are Microsoft Bing and Google Translate, with an overall BLEU score of 53.28 and 53.21, respectively.

TABLE 2
BLEU SCORE OF ARABIC-ENGLISH DATA

	CAT Tools		NMT		AI		
Type	Smartcat	Matecat	Google Translate	Microsoft Bing	Gemini	ChatGPT-4	DeepSeek
1-Gram	77.29	76.57	76.09	76.15	79.86	78.62	79.41
2-Gram	60.42	59.51	58.31	58.36	64.30	62.58	63.85
3-gram	49.46	48.42	46.70	46.74	53.80	51.79	53.15
4-gram	41.64	40.52	38.70	38.79	46.09	44.25	45.35
BLEU	55.69	54.68	53.21	53.28	59.73	57.95	59.12

As shown in Figure 4, Matecat's performance has improved for English-to-Arabic data, with an overall BLEU score of 54.68; however, it remains lower than Smartcat, which has an overall BLEU score of 55.69. Smartcat also shows a marginal advantage over Matecat in n-grams.

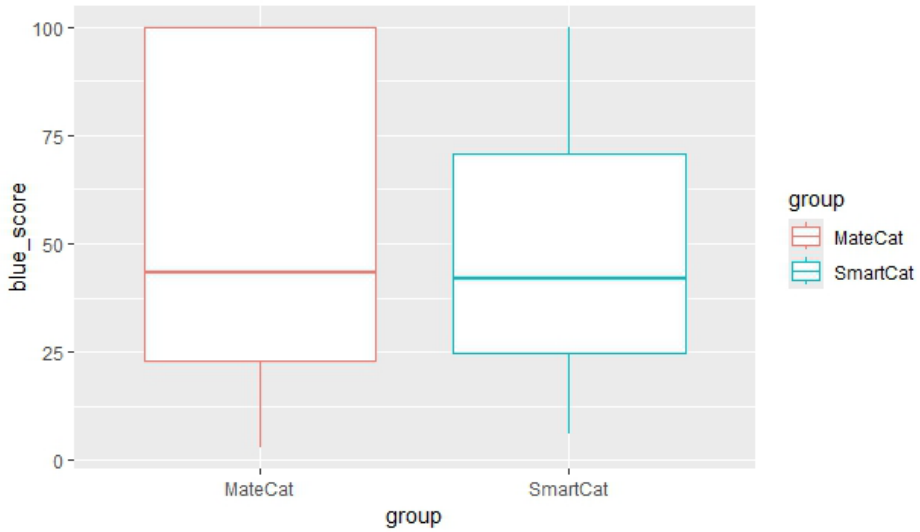

Figure 4. Boxplot of BLEU Score of Matecat and Smartcat (Arabic-English)

Figure 5 shows that Microsoft Bing Translate and Google Translate scored similar results. The earlier scored an overall BLEU of 53.21, while the latter's overall BEU score is slightly lower, at 53.28. They also show similar performance at n-gram levels.

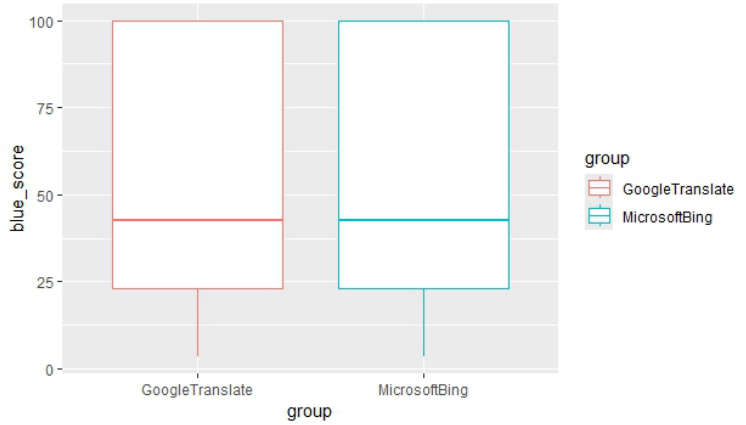

Figure 5. Boxplot of BLEU Score of Google Translate and Microsoft Bing (Arabic-English)

Figure 6 compares the BLEU score of the AI platform output. Gemini shows a relatively high BLEU score compared to DeepSeek and ChatGPT-4, with an overall BLEU score of 59.73. ChatGPT-4 has the lowest overall BLEU score of 57.95. Meanwhile, DeepSeek shows a BLEU score that falls between Gemini and ChatGPT-4, 59.12. Moreover, Gemini consistently outperforms both DeepSeek and ChatGPT-4 across all n-grams.

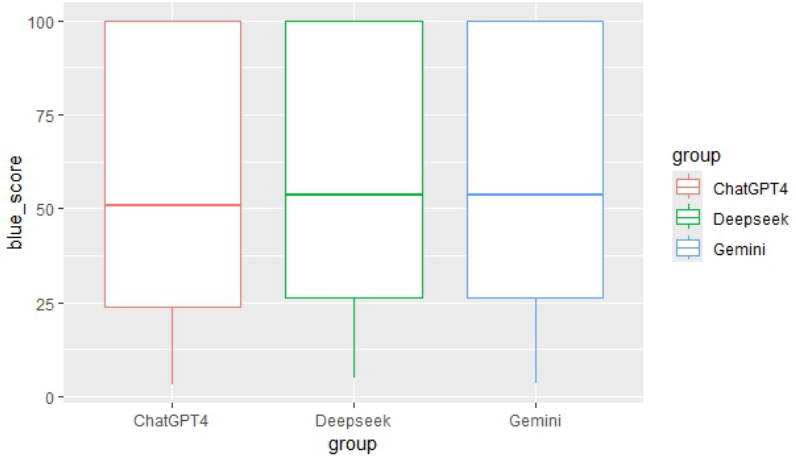

Figure 6. Boxplot of BLEU Score of ChatGPT-4, DeepSeek, and Gemini (Arabic-English)

Example 1

	BLEU	**Length ration**	**Text**
Source			Things will work out fine.
Human Translation	100.00	1.00	ستسير الأمور على ما يرام
Matecat Translation	14.06	1.00	ستعطي هذه الأشياء نتيجة جيدة
Smartcat Translation	100.00	1.00	ستسير الأمور على ما يرام

The above example shows three translations of a single English source sentence using CAT tools: Matecat and Smartcat. The length ratio indicates that both MT systems have produced a translation similar to the reference translation. However, there is a massive difference between Matecat's and Smartcat's BLEU scores. Smartcat achieved a perfect score, indicating that it produced an identical translation to the human reference translation. Meanwhile, Matecat achieved a score of 14.06. This indicates that its translation is poor and does not overlap with the human translation. The back translation, "These things will give a good result", indicates that Matecat employs a different structure and conveys a different meaning than the source and reference translations.

Example 2

	BLEU	**Length ration**	**Text**
Source			Ziri took it personally.
Human Translation	100.00	1.00	اعتبر زيري الأمر شخصيا
Matecat Translation	45.78	0.75	أخذ الأمر شخصيا
Smartcat Translation	16.23	1.50	أخذ زيري الأمر على محمل شخصي

In example 2, the length ratio indicates that the Matecat translation is shorter than the reference translation due to the omission of the proper noun "Ziri". Smartcat produced a longer sentence with less direct phrasing, "على محمل شخصي". Unlike the previous example, Matecat achieved a higher BLEU score than Smartcat. This means that its translation is closer to the reference translation, while Smartcat is less similar. To illustrate, Matecat omits "Ziri", which can be argued

to cause a loss of information and reduce the BLEU score. Additionally, the longer and shorter lengths impact the BLEU calculation.

Example 3

	BLEU	Length ration	Text
Source			Islam helped me quit drinking.
Human Translation	100.00	1.00	ساعدني الإسلام على الإقلاع عن شرب الكحول
Google Translate	19.43	0.86	ساعدني الإسلام في الإقلاع عن الشرب
Microsoft Bing	19.43	0.86	ساعدني الإسلام في الإقلاع عن الشرب

The previous example illustrates three translations of a single source sentence, translated from English into Arabic using NMT platforms: Google Translate and Microsoft Bing. The length ratio indicates that both platforms have produced translations that are slightly shorter than the reference translation. Moreover, both platforms scored a similar BLEU score, which is very low (19.43) compared to the reference translation. Correspondingly, the output of both platforms is identical, which suggests that they work similarly and depend on similar translation models. In the reference translation, the human translator adds the noun الكحول "alcohol", while Google Translate and Microsoft Bing omit it and use the general word الشرب "drinking". This, along with the shorter length ratio, contributed to the low BLEU score.

Example 4

	BLEU	Length ration	Text
Source			James gasped, clutching his chest.
Human Translation	100.00	1.00	كان جيمس يشهق ممسكا بصدره
Gemini	22.96	1.20	لهث جيمس و هو يمسك بصدره
ChatGPT-4	55.07	0.80	لهث جيمس ممسكا بصدره
DeepSeek	36.56	1.00	أخذ جيمس نفسا عميقا و هو يمسك بصدره

Example 4 illustrates three translations of a single source sentence using AI platforms: Gemini, ChatGPT-4, and DeepSeek. The length ratio indicates that Gemini's translation is slightly longer than the reference translation, while ChatGPT -4's translation is shorter. DeepSeek's translation is similar in length to the reference translation. It is worth noting that there is a discrepancy among the three AI platforms in their BLEU scores, which suggests a difference in the reference translations. ChatGPT's -4 score is the highest, 55.07, followed by DeepSeek with a 36.56 BLEU score, then Gemini with the lowest score of 22.96. Accordingly, ChatGPT-4 translation aligns closely with the human reference in meaning and structure. DeepSeek used semantic change and verbose translation of Gemini.

Example 5

	BLEU	Length ration	Text
Source			روى سامي لنا روايته لما حدث.
Human Translation	100.00	1.00	Sami told us his version of events.
Matecat Translation	68.04	1.14	Sami told us his version of what happened.
Smartcat Translation	68.04	1.14	Sami told us his version of what happened.

Example 5 shows two translations of one Arabic source sentence using CAT tools: Matecat and Smartcat. According to the length ratio, both platforms have a similar length ratio that is slightly longer than the reference translation. This is because both platforms have used "what happened" compared to the human translator's "events". Moreover, both platforms achieved a similar BLEU score, producing the exact translation, which is considered semantically accurate. Therefore, neither platform achieved a perfect BLEU score.

Example 6

	BLEU	Length ration	Text
Source			بإمكان الثعابين أن تقتل البشر.
Human Translation	100.00	1.00	Snakes can kill people
Google Translate	59.46	1.00	Snakes can kill humans
Microsoft Bing	59.46	1.00	Snakes can kill humans

Example 6 illustrates two translations of the same Arabic source sentence using NMT platforms: Google Translate and Microsoft Bing. It can be noted that the output of both platforms is consistent in terms of Length ratio and BLEU score. The length ratio is similar to the human reference; therefore, it does affect the BLEU score. On the other hand, both platforms translated the source sentence literally using the term "humans", which contradicts the reference translation "people". The difference in one word had a significant impact on the BLEU score, which relies on exact word matches.

Example 7

	BLEU	Length ration	Text
Source			لم يدع يانني شيئا لريمة.
Human Translation	100.00	1.00	Yanni did not leave Rima anything
Gemini	45.50	1.17	Yanni did not leave anything for Rima
ChatGPT-4	10.40	0.83	Yanni left nothing for Rima
DeepSeek	45.50	1.17	Yanni did not leave anything for Rima

In the above example, three translations of the same Arabic source sentence are generated using AI platforms: Gemini, ChatGPT-4, and DeepSeek. According to the length ratio, both Gemini and DeepSeek translations are slightly longer than the reference translation, resulting from different phrasing and the inclusion of the preposition. Correspondingly, the ChatGPT-4 translation is slightly shorter. Gemini achieved a BLEU score of 45.50, identical to DeepSeek's score. ChatGPT-4 performed worse, with a very low score of 10.40. Although its translation is still correct in meaning, using different grammatical structures affected the BLEU score, which favors the phrasing used in the reference translation.

V. Conclusion

This study highlighted the effectiveness of different MT platforms' translation output using the BLEU metric. The higher the BLEU score, the more similar the translation is to the human translation. By comparing the BLEU score of CAT tools (Smartcat and Matecat), NMT (Google Translate and Microsoft Bing), and LLMs (ChatGPT-4, DeepSeek, and Gemini) and analyzing 1000 sentences translated from Arabic into English and vice versa, the study revealed that LLM-based AI models surpass both CAT tools and NMT systems when translating in both directions. This finding resonates with previous studies, which highlighted the superior performance of AI-driven translation. Ghassemiazghandi's (2024) study indicated that ChatGPT-4 surpasses Matecat (CAT tool), and Alkhawaja (2024) reported the advantage of ChatGPT over Google Translate (NMT). This indicates that AI models' high advancement makes them better at dealing with the nuances of translation. Correspondingly, Gemini scores the highest BLEU score in both directions, surpassing all the other six platforms. In English-to-Arabic translation, Gemini scores an overall BLEU score of 41.30, while Google Translate's score is the lowest, at 39. 48. In Arabic-to-English, Gemini scores 59.73, while Google Translate scores 53.21. This aligns with Algobaei et al.'s (2024) study, which found that Gemini outperformed ChatGPT in addressing gender-related translation issues.

The study also reveals that these platforms struggle more with English-to-Arabic translation, which aligns with the complexity of the Arabic language in terms of morphology and diglossia. Since CAT tools tend to have the lowest scores, they can assist the human translator rather than replacing them, providing a partially automated translation. In contrast, AI models can be more helpful and preferable due to their high BLEU score. However, MT may not always be able to replicate a professional human translator's stylistic choices and conciseness, even though it may frequently produce an accurate translation that reflects the core meaning. This highlights the ongoing need for human oversight in the translation process and the limitations of current MT technologies in capturing all linguistic nuances. Further qualitative analysis of the translations in particular contexts would be needed to better understand each platform's unique advantages and disadvantages.

References

[1] Abu-Elrob, R. A., & Tawalbeh, A. I. (2025). Jordanian Facebookers' attitudes: A speech act analysis. Indonesian Journal of Applied Linguistics, 15(1), 47-58.

[2] Abdulsalami, B. A., & Akinsanya, B. J. (2017). Review of Different Approaches for Machine Translations. *International Journal of Mathematics Trends and Technology-IJMTT*, *48*(3), 197-202.

[3] Abiola, O. B., Adetunmbi, A. O., & Oguntimilehin, A. (2015). Review of the Various Approaches to Text-to-Text Machine Translations. *International Journal of Computer Applications*, *120*(18), 7-12.

[4] Adeoye, O. B. (2012). Web-Based English to Yoruba Noun-Phrases Machine Translation System. *A Thesis submitted to the Department of Computer Science*, Federal University of Technology, Akure, 2012.

[5] Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., & Ghazvininejad, M. (2022). *In-context examples selection for machine translation*. arXiv preprint arXiv:2212.02437.

[6] Algobaei, F., Alzain, E., Naji, E., & Nagi, K. A. (2025). Gender Issues between Gemini and ChatGPT: The Case of English-Arabic Translation. *World Journal of English Language*, *15*(1), 9-16.

[7] Alkhawaja, L. (2024). Unveiling the new frontier: ChatGPT-3 powered translation for Arabic-English language pairs. *Theory and Practice in Language Studies*, *14*(2), 347-357.

[8] Al-Kabi, M. N., Hailat, T. M., Al-Shawakfa, E. M., & Alsmadi, I. M. (2013). Evaluating English to Arabic machine translation using BLEU. *International Journal of Advanced Computer Science and Applications*, *4*(1), 66-73.

[9] Al-Salman, S., & Haider, A. S. (2024). Assessing the accuracy of MT and AI tools in translating humanities or social sciences Arabic research titles into English: Evidence from Google Translate, Gemini, and ChatGPT. *International Journal of Data and Network Science*, *8*(4), 2483–2498.

[10] Alsharif, B., & Khasawneh, R. (2025). Beyond the Literal: Machine Translation Performance and Stratgies in Renering Audiovisual Political Idioms. *Research Journal in Advanced Humanities*, *6*(2). https://doi.org/10.58256/4442fg06

[11] Alsharif, B., Khasawneh, R., & Alzghoul, M. (2025). Strategies of Rendering Metaphor from Arabic into English: A Comparative Study of ChatGPT and Matecat. *World Journal of English Language*, *16*(1), 45–51.

[12] Arthur, P., Neubig, G., & Nakamura, S. (2016). *Incorporating discrete translation lexicons into neural machine translation*. arXiv preprint arXiv:1606.02006.

[13] Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. arXiv preprint arXiv:1409.0473.

[14] Bédard, C. (2000). Mémoire de traduction cherche traducteur de phrases. *Traduire*, *186*(1), 41-49.

[15] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).
[16] Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., & Zou, Y. (2024). *Deepseek llm: Scaling open-source language models with longtermism*. arXiv preprint arXiv:2401.02954.
[17] Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In *Translation quality assessment: From principles to practice*, (9–38). Springer.
[18] Chatterjee, N., Johnson, A., & Krishna, M. (2007, March). Some improvements over the BLEU metric for measuring translation quality for Hindi. In *2007 International Conference on Computing: Theory and Applications (ICCTA'07)* (pp. 485–490). IEEE.
[19] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the properties of neural machine translation: Encoder-decoder approaches*. arXiv preprint arXiv:1409.1259.
[20] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., & Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1–53.
[21] Dhariya, O., Malviya, S., & Tiwary, U. S. (2017, January). A hybrid approach for Hindi-English machine translation. In *2017 International Conference on Information Networking (ICOIN)* (pp. 389-394). IEEE.
[22] Doherty, S. (2016). Translations| The impact of translation technologies on the process and product of translation. *International journal of communication*, *10*, 947–969.
[23] Ferrag, F., & Bentounsi, I. (2024). The Use of Artificial Intelligence in Academic Translation Tasks Case Study of Chat GPT, Claude and Gemini. *Ziglobitha*, (2), 173–192.
[24] Ghassemiazghandi, M. (2024). An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score. *Theory and Practice in Language Studies*, *14*(4), 985–994.
[25] Graham, Y. (2015, July). Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1804-1813).
[26] Han, L. (2016). *Machine translation evaluation resources and methods: A survey*. arXiv preprint arXiv:1605.04515.
[27] Han, L. (2022). *An overview on machine translation evaluation.* arXiv preprint arXiv:2202.11027.
[28] Hutchins, W. J. (1995). Machine translation: A brief history. In *Concise History of the Language Sciences* (pp. 431–445). Pergamon.
[29] Joshi, S. (2025). *A Comprehensive Review of DeepSeek: Performance, Architecture and Capabilities.* Retrieved from: A Comprehensive Review of DeepSeek: Performance, Architecture and Capabilities[v1] | Preprints.org
[30] Khasawneh, R., & Alsharif, B. (2025). Translating Idiomatic Expressions: A Systematic Review. Forum for Linguistic Studies, 7(10). 881-893.
[31] Karpina, O. (2024). Comparative study of modern CAT tools: Smartcat vs Matecat. In *Пріоритети германської і романської філології*. Волинський національний університет імені Лесі Українки.
[32] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 1-13.
[33] Khasawneh, R. R., Moindjie, M. A., & Kasuma, S. A. A. (2025). Diachronic Translation of Figures of Speech in Antara's Mu'allaqa. *World Journal of English Language*, *15*(3), 290-290. https://doi.org/10.5430/wjel.v15n3p290
[34] Kleinman, G. (2023). Demystifying the BLEU Metric: A Comprehensive Guide to Machine Translation Evaluation | traceloop Blog. *Demystifying the BLEU Metric: A Comprehensive Guide to Machine Translation Evaluation | Traceloop - LLM Application Observability*.
[35] Maučec, M. S., & Donaj, G. (2019). Machine translation and the evaluation of its quality. In *Recent trends in computational intelligence*. IntechOpen.
[36] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). *Large Language Models: A Survey.* arXiv preprint arXiv:2402.06196
[37] Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). *Adaptive machine translation with large language models.* arXiv preprint arXiv:2301.13294.
[38] Mukhtar, I. A. (2025). Translation and Technology. *Transcultural Journal of Humanities and Social Sciences*, *6*(2), 269–283.
[39] Pan, R., Ibrahimzada, A. R., Krishna, R., Sankar, D., Wassi, L. P., Merler, M., & Jabbarvand, R. (2024, April). Lost in translation: A study of bugs introduced by large language models while translating code. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (pp. 1–13).
[40] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
[41] Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., & Tao, D. (2023). *Towards making the most of ChatGPT for machine translation*. arXiv preprint arXiv:2303.13780.
[42] Poibeau, T. (2017). *Machine translation*. MIT Press.
[43] Raymond, D. (2024). Top 10 Cons & Disadvantages of Google Translate | Project Managers Blog. *2025. Top 10 Cons & Disadvantages of Google Translate.*
[44] Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, *56*(2), 593-619.
[45] Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., & Ryder, N. (2022). ChatGPT: Optimizing Language Models for Dialogue. *OpenAI blog*, *2*(4), 9-27.
[46] Sharma, N., Bhatia, P. G., & Singh, V. G. (2011). *English to Hindi statistical machine translation system* [Doctoral dissertation]. Thapar University.
[47] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., & Blanco, L. (2023). *Gemini: a family of highly capable multimodal models*. arXiv preprint arXiv:2312.11805.

[48] Wołk, K., & Marasek, K. (2015). Neural-based machine translation for medical text domain. Based on European Medicines agency leaflet texts. *Procedia Computer Science*, *64*, 2-9.
[49] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., & Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation.* arXiv preprint arXiv:1609.08144.
[50] Yang, Z., Liu, F., Yu, Z., Keung, J. W., Li, J., Liu, S., & Li, G. (2024). Exploring and unleashing the power of large language models in automated code translation. *Proceedings of the ACM on Software Engineering*, *1*(FSE), 1585-1608.

Razan R. Khasawneh is a part-time lecturer in the Department of English Language and Translation at Amman Arab University, Jordan. Her research interests include Diachronic Translation, Literary Translation, and Machine Translation.

Bilal I. Alsharif is an Assistant Professor in the Department of Applied Linguistics at Al-Zaytoonah University of Jordan. His research interests include Laboratory Phonology, Sociophonetics, Arabic dialectology, and Machine Translation.