Research Journal in Advanced Humanities

RJAH

https://doi.org/10.58256/vkyfk485

Open Access

Check for updates

RESEARCH ARTICLE

Section: *Literature, Linguistics & Criticism*

Assessing the effectiveness of voice-translation tools for translating medical information from English to Arabic

Zakaryia Almahasees[1]*, Abdelkarim Allababneh[2], Bassam Saideen[3], Noor Darwish[4], Bilal Alderbashi[3], Majed Yahya Abdulkarim[5] & Islam Husienat[6]

[1]Department of Translation, Yarmouk University, Jordan

[2]Department of English and Translation, Sultan Qaboos University, Oman

[3]Department of English/ Translation, Faculty of Arts, Isra University, Jordan

[4]Department of Basic Sciences, Humanities, Faculty of Arts and Humanities, Applied Science Private University, Jordan

[5]Department of English and Literature, Faculty of Arts, Isra University, Jordan

[6]King Abdullah University Hospital, Jordan

*Correspondence: z.almahasees@yu.edu.jo

ABSTRACT

The study aims to assess the performance of Gemini, ChatGPT, and DeepSeek models in translating medical information from English to Arabic. Using a comparative descriptive design, the current study tests 40 English audio clips of medical information and evaluates their 120 Arabic translations using the Multidimensional Quality Metrics (MQM). The performance of each model was evaluated along two dimensions of translation quality: accuracy and fluency, including medical terminology, semantics, grammar, and clinical instructions. The comparative analysis revealed accuracy and fluency problems, including problems related to the accuracy of medical terms, omissions, clinical plausibility, and emergency communication. In the accuracy assessment, ChatGPT scored 18, Gemini 14, and DeepSeek 10 points. The current study stresses the importance of differentiating between linguistic fluency and medical accuracy and the significance of human supervision.

KEYWORDS: medical translation, artificial intelligence, ChatGPT, Gemini, DeepSeek, English–Arabic Translation, Multidimensional Quality Metrics (MQM)

Research Journal in Advanced Humanities

Volume 7, Issue 4, 2026

ISSN: 2708-5945 (Print)

ISSN: 2708-5953 (Online)

ARTICLE HISTORY

Submitted **27 August 2026**

Accepted: **14 October 2026**

Published: 08 **November 2026**

HOW TO CITE

Almahasees, Z., Allababneh, A., Saideen, B., Darwish, N., Alderbashi, B., Yahya Abdulkarim, M., & Husienat, I. (2026). Assessing the effectiveness of voice-translation tools for translating medical information from English to Arabic. *Research Journal in Advanced Humanities*, 7(4). https://doi.org/10.58256/vkyfk485

Published in Nairobi, Kenya by Royallite Global, an imprint of Royallite Publishers Limited

© 2026 The Author(s). This is an open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

The fast development of AI has greatly affected the translation industry through developments in NMT, ASR, NLP, and LLM. These innovations expand the possibilities for multilingual communication by providing advanced language processing in both text and speech forms (Hartung & Kleinstreuer, 2025). Generative AI tools are now used routinely for language-related tasks, from feedback on learners' English (Haider, 2026a) to the translation of specialized texts. Currently, the translation system is not confined to the text-to-text translation task, but incorporates such components as ASR, speech-to-text translation, and speech synthesis into the multilingual communication framework. Large-scale multilingual translation models demonstrate the ever-expanding possibilities of AI systems to process linguistic data in different languages and modalities. Thus, Costa-jussà et al. (2022) explored different approaches to increasing the coverage of machine translation across languages. Also, Barrault et al. (2023) developed SeamlessM4T, which is a multilingual system that covers several tasks related to speech and text translations. Such innovations provide numerous possibilities for using AI translation solutions in various areas, including medicine, where multilingual communication is becoming an increasingly important issue.

Medical translation can be considered one of the most challenging areas for machine translation owing to the need to communicate highly specific knowledge which might affect the understanding of the patients, their decisions concerning treatment and health outcomes (Khujakulov, 2026). Medical texts involve specialized terms, diagnostic classifications, instructions for medicines, numbers, clinical advice, and notions of uncertainty. This means that the quality of the translation in the area under consideration should not only reflect grammatical and stylistic proficiency but also convey the medical meaning of the text accurately. Indeed, misstatement of the medicine dose, loss of a negative phrase or replacement of a specialized diagnosis can change the medical meaning of the message (Mahdi & Mohammed, 2025). In the article about risks and prospects of machine translation in the medical and legal fields, Vieira et al. (2021) pointed out dangers related to the irresponsible application of machine translation in crucial situations. Similarly, the World Health Organization (2024) focused on the importance of accuracy, reliability, safety, and proper governance of AI systems in the health field.

The importance of reliable medical translations is particularly relevant in multilingual healthcare systems, where communication difficulties resulting from language barriers can affect the interaction between the physician and the patient (Almahasees & Almahasees, 2025). Patients need accurate information regarding their medical condition, prescribed treatments, the schedule of medications, possible side effects, and other follow-up information. When the health-related information is provided in a language that is incomprehensible to the patient, the communication difficulties can prevent the patient from making an informed decision. Machine-based translation systems can help solve the problem of fast access to information in various languages, especially when professional medical interpreters are unavailable. Nevertheless, the availability does not necessarily mean reliability in the context of clinical communications. Vieira et al. (2021) noted that social implications of machine translations need to be compared to the risks associated with translation errors. The relevance of such research is emphasized by Carreras Tartak et al. (2025), who explored machine-translated discharge instructions in the emergency department.

The use of voice-translation technologies brings with it challenges in the translation process. Unlike the translation of texts, voice translation involves translating spoken words, which means that the technology involved should be able to understand the speech, recognize the structures, and generate a translated version of the message. Some systems use a combination of automatic speech recognition and translation while others can perform speech translation on their own. There is always the risk of errors during speech recognition, interpretation of meaning, and generation of a translated message (Almahasees & Husienat, 2026). Radford et al. (2023) made remarkable progress in multilingual speech recognition using large-scale weak supervision techniques while Barrault et al. (2023) showed the effectiveness of using speech and text translation in one system. Nonetheless, the advancement in speech processing performance does not guarantee voice translation accuracy in medical scenarios.

Voice translation in the medical field has its peculiarities because often there can be a lot of specialized terminology, nomenclature of pharmaceuticals, anatomical terms, abbreviations, measurements made in laboratory tests, as well as instructions that require accurate quantification. Besides, this kind of information may have different delivery rates, variations in pronunciation, unfamiliar accents, hesitations, and background

noise. This may influence speech recognition, and consequently, the accuracy of translation (Sapawi & Yusoff, 2026). For example, incorrect recognition of the name of some pharmaceutical can result in incorrect translation, although the process of translation itself is done correctly. On the other hand, while the system can correctly recognize the English expression, it may provide an inappropriate Arabic medical term. Therefore, it is necessary to differentiate between recognition errors and translation errors when assessing voice translation systems. The technological improvements mentioned by Radford et al. (2023) and Barrault et al. (2023) can serve as a base for solving the problem; nevertheless, a specialized evaluation is still needed.

The use of English and Arabic entails further complications due to significant dissimilarities in linguistic systems and terminology of the languages involved. English is commonly used in medical research, medical records, pharmaceutical information, and global health care communication; Arabic is the main means of communication for large groups of people in the Middle East and North African regions. It follows that the translation of information in the medical field from English into Arabic will have to take into account differences in morphology, syntax, semantics, and terminology. Some medical expressions of the English language have appropriate Arabic counterparts, some can be translated descriptively, some can be transliterated (Haider, Saed, et al., 2025), and some will have to vary in accordance with the target audience. Besides, the terminology in medical communication will depend on the level of specialization of the target audience (medical practitioners or patients). These facts show that the quality of English–Arabic medical translation can be determined not only through terminology accuracy but also through communicative suitability (Boulesnam & Boucetti, 2025).

Recent developments in generative AI have further complicated the relationship between translation accuracy and linguistic fluency. Large language models can produce coherent, grammatically acceptable, and contextually plausible translations, often with a degree of linguistic flexibility that distinguishes them from earlier machine translation systems. However, fluent language generation does not necessarily guarantee faithful preservation of the source message. In medical contexts, a system may replace a specialized expression with a broader term, omit an expression of uncertainty, or reformulate an urgent instruction in a manner that reduces its original force (Trigka & Dritsas, 2025). Such changes may produce natural Arabic sentences while altering medically relevant information. The importance of systematic error-based evaluation is supported by Freitag et al. (2021), who demonstrated the value of expert evaluation grounded in explicit error analysis. Their work highlights the limitations of relying exclusively on overall quality judgments when assessing modern machine translation systems.

The current study, titled "Assessing the Effectiveness of Voice-Translation Tools for Translating Medical Information from English to Arabic," examines the effectiveness of Gemini, ChatGPT, and DeepSeek in translating medical information from English to Arabic through speech translation. The method used in the research design entails the selection of a purposive corpus consisting of 40 English medical audio excerpts obtained from freely accessible YouTube videos. These audio excerpts include information about medical conditions, symptoms, directions on medications, tests and diagnosis, and clinical advice. Translations were evaluated based on the accuracy and fluency criteria in MQM with regard to mistranslation, omission, addition, and language errors. A qualitative analysis of representative examples was also used to complement the frequency-based approach through which translation practices that could lead to the retention of the medical meaning in the translation process can be identified.

The research makes contributions to the following three interconnected fields: AI-powered translation technology, medical communications, and the field of English-Arabic translation studies. First, it expands the assessment of generative AI translation systems to include the translation of spoken medical texts, which is an undertaking that involves technical and linguistic complexities. Second, it explores the language pair of English and Arabic where terminology and structural differences should be taken into account. Third, it applies MQM-based error classification to distinguish between semantic errors and linguistic errors, thus providing a more refined evaluation of the translation process. Overall, the study seeks to contribute to the sensible use of AI-powered translation technologies in healthcare communications.

2. Literature Review

2.1. Artificial Intelligence and the Transformation of Specialized Translation

The development of neural machine translation and large language models has changed the technology behind

current translation methods. In the past, statistical machine translation models primarily focused on the probability distribution of source language and target language sequences, whereas neural models focus on distributed representation and context-based translation (Almahasees, 2023). Later, multilingual and generative language models expanded the process of translation from sentence-level to task-based translation, incorporating instruction following and contextual interpretation. As evidenced by the No Language Left Behind project introduced by Costa-jussà et al. (2022), multilingual language models have enabled the expansion of translation to those languages that have been previously neglected technologically. Moreover, research into large language models has highlighted the increasing ability of generative models to undertake multilingual tasks despite not being engineered specifically for such tasks.

While there has been development in the area of general language generation, it does not mean that these advancements automatically contribute to improved levels of translation reliability. Specialized translation requires maintaining certain domain-specific concept relations, terminology, and communicative functions that are not always reflected in the criteria used in evaluations focused on general fluency. In medicine, for example, using another semantically similar but clinically different term in place of an original one can produce a very fluent translation with poor information. It is even more important in the case of generative AI models that produce linguistically correct but not strictly equivalent translations. According to Freitag et al. (2021), error-focused evaluation conducted by experts is particularly useful when evaluating good machine translation because of possible differences among various translations.

Another factor has to do with the balance between flexibility in generation and faithfulness of translation. Language models may modify vocabulary and syntax and even explain concepts to make the text more understandable for certain target audiences. Although this can be useful in facilitating translation, it can lead to unforeseen semantic shifts where translation must adhere closely to the source text. This creates a dilemma between clarity and precision especially in communication (Karanikolas et al., 2023). Therefore, the assessment of the ability of generative AI in medical translation must consider a criterion that can distinguish acceptable deviations from those that change the meaning of the medical text.

2.2. Advances and Limitations in Automatic Speech Recognition and Speech Translation

Speech translation is set apart from text translation through its requirement for the analysis of acoustic input before and/or simultaneously with cross-language translation (Mingote et al., 2023). Conventional cascaded architectures often employ automatic speech recognition and machine translation, but more modern methods have explored direct and integrated models of speech-to-text and speech-to-speech translation. Radford et al. (2023) found that large-scale weakly-supervised learning has a major impact on improving speech recognition in various acoustic and linguistic settings. The Whisper initiative represents a breakthrough in the field of multilingual speech processing especially in light of the fact that good speech recognition is a necessary prerequisite for voice translation tasks. Nonetheless, high-quality general speech recognition does not necessarily mean quality medical speech recognition.

More progress has been made by Barrault et al. (2023), who have developed a framework known as SeamlessM4T. SeamlessM4T is a multilingual and multimodal framework that is able to translate both speech and text. This work represents an example of a larger trend toward more integrated frameworks that are able to process multiple modalities of languages using one architecture. These achievements are particularly relevant for healthcare communication since they suggest that there is potential for multilingual communication that is not associated with separate transcription and translation processes. However, the existence of integrated speech processing capabilities does not eliminate the need for domain-specific testing.

The difference between errors of speech recognition and translation errors is key when evaluating the quality of voice-translation technology. When using the cascaded approach, an error made in one step may carry forward to the next stage, resulting in a clinically wrong but linguistically correct sentence in the target language. On the other hand, the integrated system would have no transcript stage, making it difficult to discern whether the error resulted from the acoustic processing or the cross-linguistic generation. A mistake in recognizing the numerical dose could lead to an error in the output, despite the clinical Arabic sentence being correctly constructed. Ideally, research on medical voice translation would retain the source audio and get a correct transcript of the source language. This would make it possible to evaluate both the transcription and

the translation against the relevant reference. Without these controls, errors in the output could be categorized, although their technical source would remain unknown (Kuhn et al., 2024).

These methodological issues reveal a limitation of applying speech translation research to healthcare. While general benchmarks of multilingual speech recognition usually focus on overall recognition or translation accuracy, the medical communication process requires that special attention be paid to possible errors of low probability, which may still carry much importance. Therefore, even though a system performs well with regular speech, it can still incorrectly recognize the name of the drug, overlook negation, or adjust a numerical instruction.

2.3. Machine Translation in Healthcare: Accuracy, Safety, and Clinical Communication

Incorporation of machine translation in the health sector has received increasing attention from researchers due to the technology's promise to overcome the problem of language barriers and improve access to medical information. Machine translation provides instant translation in multilingual scenarios involving medical education, communications, and general health information. However, the benefits that may accrue from the use of such technologies need to be considered alongside the potential impacts that may arise from erroneous or partial translation (Karakus et al., 2025). In a review of literature on the societal impact of machine translation in medical and legal fields, Vieira et al. (2021) highlight the need for comprehension of translation technologies within their intended context of use.

The research on machine-translated patient instructions provides a rather relevant perspective. Carreras Tartak et al. (2025) analyzed the effectiveness of Google Translate in translating discharge instructions in different languages. The relevance of this study is determined by the fact that patient instructions typically contain information about medications, warning signs, the need for further actions, and the circumstances under which patients must visit the doctor again. In contrast to other kinds of texts, patient instructions have an immediate function of providing guidelines for patients' actions. Thus, a translation that maintains the topic but changes the instructions can be ineffective.

One of the crucial issues that need to be raised in the context of translation in the healthcare sector is the insufficiency of the equivalence between linguistic understanding and medical correctness of the translation. A translated phrase might be easily comprehensible for the reader but would deliver an incorrect dose, warning, or description of the illness. This problem is particularly acute in relation to AI-powered translation due to the fact that contemporary AI systems tend to generate linguistically correct sentences. Therefore, such a sentence might seem trustworthy despite being medically incorrect. The World Health Organization (2024) states that healthcare-related large multimodal AI systems require particular attention in terms of factual trustworthiness and safety (see also Almahasees & Husienat, 2024).

The scholarly literature clearly shows that any translation errors must be judged based on the consequences they could have rather than being perceived as equally important linguistic issues. Even a simple error in punctuation is unlikely to affect the understanding of the patient, while failure to include the negative form of a phrase or the inclusion of an incorrect dosage could change the meaning. This means that it would be reasonable to classify translation errors by type and, where necessary, evaluate their severity. However, it is extremely important to differentiate linguistic error detection from the process of clinical risk assessment. The fact that the error in translation is detected does not mean that it will cause patient harm (Yelahina & Fedchyshyn, 2025).

Taken together, there are indications from healthcare translation studies that machine translation is valuable but requires assessment according to criteria that pertain to medicine. Also, it becomes evident that the performance of machine translation as a whole cannot be assumed to determine accuracy in specific clinical genres. However, much of the available literature deals with patient instructions written down or machine translation services generally. In the case of the translation of oral communication in medicine using generative artificial intelligence, there are other challenges that arise.

2.4. English–Arabic Medical Translation and Terminological Challenges

Translation of medical terminology from English to Arabic requires particular attention to the relationship between specialized terminology, linguistics, and communicative purposes. Despite the fact that Arabic has

many terms used in medicine, the choice of proper equivalents is made difficult by such factors as different lexicalization, specialization, and addressees of a text. For example, one medical term may have a technical equivalent in Arabic, which is to be used in medical records, and an explanatory variant which is to be used in patient communication. Both terms may refer to similar concepts but can be interchanged in a particular medical context only. In the case of AI-based translation, the risk of generalization of medical terminology arises because of the choice of a more general term in a target language to replace a specialized one in the source language (Almahasees & Husienat, 2024). Similar terminological difficulties have been documented in other specialized domains, including the machine and AI translation of English legal, business, and financial acronyms into Arabic (Haider, Abu Akar, et al., 2025; Haider & AlKhatib, 2024).

Morphology and syntax of Arabic pose some challenges in automated translation. The morphology of the Arabic language is highly divergent from that of English in many aspects including structure, agreement rules, sentence construction and grammatical structures. This divergence makes it necessary for the translation process to generate translations that are both grammatically and semantically correct. However, the mere fact that the Arabic translation generated is grammatically correct does not mean that it is terminologically equivalent. For example, 'renal impairment' and 'renal failure' refer to two different medical terms although they relate to each other. Also, 'liver enzymes' and 'liver function' are different descriptors in medical tests (Delfani et al., 2024).
Sources of research material for English-Arabic medical translation are relevant to the development of more systematic evaluation processes. Bilingual corpora can help in identifying technical terms, collocation patterns, and translation norms, and also provide data for testing machine-translated texts. Corpus methods also make the analysis of Arabic and English texts more systematic and less exposed to researcher bias (Haider, 2019a, 2019b, 2019c). Nevertheless, a medical corpus in written form does not cover all the difficulties that arise during spoken medical interactions such as acoustic confusion, pronunciation differences, and speech recognition errors. Therefore, it is necessary to consider the existence of sources in written form as a supplement to, rather than a replacement for, English-Arabic medical voice translation research (Delfani et al., 2024).

2.5. Translation Quality Assessment and the MQM Framework

The assessment of the quality of machine translation has seen considerable methodological advancements due to the increasing capabilities of machine translators to create fluent texts in the target language. The traditional evaluation techniques were mainly based either on automatic measures using reference corpora or on holistic assessments made by humans. While those techniques may indeed yield useful results about the general performance of the system, they might not suffice for locating specific errors that influence the meaning of the text (Ismayilli, 2024).

Freitag et al. (2021) carried out a massive-scale study on human evaluation of machine translation based on the Multidimensional Quality Metrics approach. In their study, the authors used professional translators and annotated errors for assessing the performance of high-quality machine translation systems. The key point is that system rankings based on this kind of assessment are different from the rankings derived using crowd-based evaluations, meaning that the evaluation approach can actually affect the results of assessments.

MQM (Lommel et al., 2024) can easily be applied to medical translation due to its ability to categorize translation mistakes based on their linguistic and semantic features. In terms of the accuracy dimension, mistranslation is a mistake wherein the target translation includes information which is not included in the source; omission is the lack of information in the source language. Addition is the insertion of information which is not sufficiently justified by the source text; untranslated text is the text which has not been adequately translated. Meanwhile, the fluency dimension includes the mistakes associated with the proper use of the target language, including grammar, spelling, punctuation, and word order (Li, 2026).

The distinction between accuracy and fluency becomes especially relevant in the case of medical voice translation. An English sentence that suggests that a certain medicine could have a specific side effect will lose the word that shows that this side effect is possible when translated into Arabic. As a result, this Arabic sentence can be correct grammatically, yet the content of the medical statement will change due to the loss of certainty. In a similar way, the translation of the phrase stating that one should immediately call the emergency department can be generalized to seeking medical help without mentioning the necessity of doing it urgently (Muhanna, 2025).

Another aspect to consider in terms of methodology relates to the level of seriousness of translation mistakes. While using MQM will help classify mistakes according to type and severity, it is important to note that the number of translation mistakes does not necessarily imply clinical significance. A grammatical mistake on the part of the translator and a wrong dose of the medicine are not necessarily equal simply because both are mistakes. In a study involving translation of medical texts, it would be helpful to separate the seriousness of the language mistakes from clinical risks. Clinical risk, however, requires certain medical knowledge to determine. For the current research project, the accuracy and fluency components of MQM constitute a suitable framework to evaluate the outputs produced by Gemini, ChatGPT and DeepSeek in Arabic. In the former case, it is possible to compare the outputs in terms of terminology usage, instructions, negations, uncertainty, and other issues that are important from the perspective of medicine. Meanwhile, in the latter case, it is possible to measure grammatical and linguistic correctness. The combination of statistical analysis with the analysis of representative examples explains the reasons behind certain translation mistakes. This methodology is consistent with the general concept provided by Freitag et al. (2021).

There are several important developments in multilingual machine translation, automatic speech recognition, and quality assessment of translations found in the reviewed literature. Researching big multilingual models has expanded the technological opportunities for communication across languages, whereas developments in speech processing technology systems have increased the possibility of implementing spoken language translation. Studies focused on the use of machine translation in the healthcare field demonstrate the real opportunities of such technologies, as well as the need to evaluate the accuracy of information delivered to patients. At the same time, research on MQM evaluation shows the importance of detailed error marking (Mejaški, 2026).

However, some aspects are still not fully explored by this line of research. First, the performance of multilingual speech processing systems on the general benchmark does not guarantee their performance on medical speech. Second, the findings related to the machine translation of written instructions are not applicable to oral communication because there are other ways in which errors can occur due to acoustic processing. Third, a general benchmark may not indicate the performance of large language models on English-Arabic terminology and instructions. Fourth, since generative AI models become more and more fluent, it is essential to have a way to evaluate the changes in the meaning.

This paper aims to examine the translation of medical discourse from English into Arabic utilizing Gemini, ChatGPT, and DeepSeek systems. The contribution of this research can be found in analyzing the translated Arabic versions of the texts on the basis of a unified MQM standard, paying special attention to the accuracy and fluency of the translations. In addition, the objective of the paper is to detect the common mistakes related to the medical terminology, clinical instructions, numbers and semantics in order to differentiate them from the grammatical mistakes that occur in the target language.

3. Methodology

3.1. Research Design

This study makes use of the comparative research design, which combines quantitative error analysis with qualitative investigation in order to measure the effectiveness of Gemini, ChatGPT and DeepSeek systems in translating spoken medical data from English to Arabic. The methodology, which is based on Multidimensional Quality Metrics (MQM), provides a structured way to identify and categorize the translation errors (Freitag et al., 2021). The error analysis focuses on two main characteristics of translation quality, namely, accuracy, which means preservation of the source language meaning and information, and fluency, which means linguistic correctness of the Arabic translation. The comparative design allows for an analysis of the nature, number, and distribution of errors produced by the systems. Quantitative techniques have been used to calculate the frequency and percentage of errors, and qualitative techniques have been used to analyze some cases in order to understand their medical and linguistic significance.

3.2. Data Collection and Corpus Selection

The data set comprises 40 English medical audio samples which were selected from public videos on the website YouTube. Purposive sampling has been used for selecting the samples in order to ensure that the samples

selected would cover different forms of medical information and contain the linguistic characteristics needed for evaluating translation accuracy. The samples include information about medical diagnoses, symptoms, drug prescriptions, treatment processes, medical tests, and warnings. The focus was on those samples that contain specialized vocabulary, numbers, language of possibility, negation, and instructions. Each sample is considered a separate unit of analysis. The English speech has been transcribed in order to have the source-language version against which the translations into Arabic can be compared. The data set has been limited to the translation from English to Arabic only.

3.3. **Translation Tools and Experimental Procedure**

The current study compared three artificial intelligence generation models—Gemini, ChatGPT, and DeepSeek—regarding their language-processing capacity and suitability for use in AI-powered medical translation. Forty English medical audio clips were translated into Arabic by each model, generating 120 outputs in total. For this purpose, the same source material and the same translation criteria were used for all three AI systems. According to the criteria, the system had to translate the information in the audio files into Arabic without summarizing, omitting, or adding anything extra to the content of the text. Every excerpt was done separately so as to avoid possible interference from previous tasks. The Arabic translations were collected and recorded verbatim as the actual outputs of each system, and put together with the original English texts for further comparison through the MQM tool.

Objective of the Experiment

The objective of this experiment was to test the accuracy and fluency of Arabic translations provided by the models of Gemini, ChatGPT, and DeepSeek. The former quality was estimated through the presence of preserved medical terminology, diagnosis data, prescriptions, numeric data, and relevant clinical phrases, whereas the latter quality was estimated through grammaticality, word order, orthographic correctness, and a generally linguistically acceptable style of all translations. For consistency, all translations were compared to the same English source transcripts. In this experiment, no attention was paid to either the processing speed or the computational efficiency of the different models; instead, an attempt was made to find out the difference in error rates among the three models.

3.4. **Translation Quality Assessment Framework**

The current study adopts the MQM approach for evaluating the quality of the medical translations performed using Gemini, ChatGPT, and DeepSeek. Specifically, MQM is chosen because of its systematic approach to identifying and classifying translation errors based on linguistic and semantic grounds (Freitag et al., 2021). The analysis focuses on two main dimensions, namely Accuracy and Fluency. Accuracy measures the level to which Arabic translations retain the meaning of the original medical texts, including the use of specialized terminology, information about diagnosis, instructions on medications, numbers, and clinical warnings. Errors in this dimension include mistranslation, omission, addition, and untranslated text. Fluency measures the linguistic quality of Arabic outputs in terms of grammar, word order, spelling, and punctuation. Both the original and translated texts are analyzed in comparison with each other, and the translation errors are identified and classified based on specific criteria.

3.5. **Data Analysis**

The research makes use of quantitative and qualitative approaches in order to compare the translation performance of Gemini, ChatGPT, and DeepSeek. In the quantitative section, the errors in the Accuracy and Fluency dimensions of the MQM are identified and categorized, and their frequencies and percentages are calculated among the 120 translated texts. The frequency of each category of errors is computed separately for each system in order to identify the recurrent errors and differences in performance among the systems. The qualitative approach entails the selection of ten representative medical cases among the 40 excerpts in the source language for comparison. Each of these cases is presented in tabular form, which presents the English source text and the Arabic translation produced by each of the three systems, followed by an analysis in terms of terminology, meaning, grammar, and clinical information.

4. Analysis and Findings

This section presents a comparative analysis of the medical translations produced in Arabic by Gemini, ChatGPT, and DeepSeek using the MQM criteria. In particular, the analysis focuses on two major dimensions: Accuracy, which implies the retention of medical meanings, terminology, and clinical instructions, and Fluency, which refers to the correctness of the grammar, word order, and language quality. The experimental design entails 40 English medical audio samples and 120 corresponding Arabic translations, where 40 samples per model are used. Ten representative cases are chosen for a detailed qualitative analysis in order to establish translation patterns and reveal possible differences in the way specialized medical information is translated by the systems. Every sample is provided in a comparative table and then analyzed for accuracy and fluency. All Arabic translations reported in the tables are the actual outputs generated by Gemini, ChatGPT, and DeepSeek for the source audio clips, reproduced verbatim without any editing. Special attention is paid to errors related to medical terminology, probabilistic statements, negations, medications, and clinical urgency, which are key components in ensuring the communicative purpose of medical information.

4.1. Qualitative Comparative Analysis

Example 1: Acute Myocardial Infarction (AMI)

In the first example, the translation of a cardiovascular diagnosis is considered, with particular attention to medical terminology and the difference between technical and common terms.

English source text	Gemini	ChatGPT	DeepSeek
The patient was diagnosed with acute myocardial infarction.	شُخِّص المريض باحتشاء عضلة القلب الحاد.	تم تشخيص المريض باحتشاء حاد في عضلة القلب.	شُخِّص المريض بنوبة قلبية حادة.

The provided translations demonstrate variability in terminological exactness. In the translations produced by Gemini and ChatGPT, the medical term احتشاء عضلة القلب is used, which translates into myocardial infarction. The diagnostic name and the adjective acute remain in all translations, although there is some discrepancy in the syntax. On the contrary, the translation made by DeepSeek, نوبة قلبية حادة, uses a more common term that can be translated as "heart attack." From the perspective of terminology exactness, this term is less exact than the previous two. According to MQM, such ambiguity is not a mistranslation since this term may be relevant for patients. Still, when medical terminology is needed, the previous terms may be more appropriate.

Indeed, all three translations are grammatically correct and meaningful in the target language. Gemini offers a precise sentence which can be used in the clinical context, while ChatGPT generates a sentence with a passive voice construction, which is acceptable in the medical literature as well. Finally, DeepSeek offers a smooth translation, easily understandable by common people. There were no errors identified regarding the grammar and spelling of the translations. Hence, it is clear that differences in medical terminology may exist without any problems with translation fluency.

Example 2: Medication Dosage and Administration Frequency

The current paper investigates the translation process of instructions for medicines, paying special attention to the accuracy of transferring numerical information about dosages, frequency, and timing.

English source text	Gemini	ChatGPT	DeepSeek
Take 5 mg twice daily after meals.	تناول 5 ملغ مرتين يوميًا بعد الوجبات.	تناول جرعة مقدارها 5 ملغ مرتين يوميًا بعد تناول الطعام.	تناول جرعة 5 ملغ مرتين يوميًا بعد الطعام.

All three translations contain all key aspects of the drug instructions. The dose amount of 5 mg is mentioned correctly in the translation, twice-a-day intake continues to be required, and the instruction to take the drug after meals is kept intact. In this regard, differences between the translations are mainly of a stylistic nature and not semantically oriented. ChatGPT includes the phrase جرعة مقدارها, while DeepSeek adds just جرعة. Both phrases

fit the context well and explain what kind of dose is being referenced in the translation without changing the required dose. Gemini provides a slightly briefer version of the translation that does not distort the original source. In MQM standards, such differences should not be considered accuracy errors since they do not change the drug instructions.

The three translated sentences display acceptable Arabic grammar and proper word order. The translation done by Gemini is brief and clear, fitting well into the terse language that is characteristic of pharmaceutical instructions. ChatGPT's translation is longer but still fits the language criteria, while DeepSeek managed to find the middle point between brevity and clarity. None of the three translations contains an apparent mistake related to either grammar or orthography. Thus, in this case, we observe how several Arabic translations can be equally accurate and fluent.

Example 3: Medical History of Diabetes Mellitus

This paper addresses the issues surrounding the translation of negative medical history by paying special attention to the differences between the lack of a recorded history of the disease and the lack of a diagnosis.

English source text	Gemini	ChatGPT	DeepSeek
The patient has no history of diabetes mellitus.	ليس لدى المريض تاريخ مرضي للإصابة بداء السكري.	لا يوجد لدى المريض تاريخ مرضي للإصابة بداء السكري.	المريض ليس مصابًا بداء السكري.

Both Gemini and ChatGPT keep the core sense of the English text through an explicit mention of the lack of a medical history of diabetes mellitus. It should be noted that the concept of medical history is preserved in both translations, which is critical for the original statement. However, DeepSeek provides an example of a translation that reformulates the statement, implying that the patient has no diabetes. It might seem that the two statements are very similar; however, their meanings are different since a statement about the lack of medical history does not mean that the patient has a particular diagnosis now. According to the MQM standard, this change can be regarded as a possible mistranslation.

The three translations are all grammatically sound and natural in Arabic. Gemini and ChatGPT have used the appropriate formal language for a patient's medical file, but DeepSeek has used a more informal structure. However, DeepSeek still uses fluent language even though there is a difference in the semantics, as pointed out in the Accuracy section above. This becomes significant as the readers might understand DeepSeek as a clear translation without realizing that the medical information it provides is different from that originally written.

Example 4: Adverse Drug Reactions and Clinical Probability

This example looks into the translation of adverse effects of drugs with emphasis on maintaining modalities indicating clinical possibility.

English source text	Gemini	ChatGPT	DeepSeek
This medication may cause dizziness and drowsiness.	قد يسبب هذا الدواء الدوار والنعاس.	قد يؤدي تناول هذا الدواء إلى الشعور بالدوار والنعاس.	يسبب هذا الدواء الدوار والنعاس.

The following three examples of translations show how the degree of certainty conveyed in the source statement varies in different versions of the translation. The translations provided by Gemini and ChatGPT correctly convey the modal verb may through the Arabic particle قد, indicating the possibility rather than the certainty of adverse reactions. ChatGPT also adds an explanation using the phrase تناول هذا الدواء without modifying the original meaning. However, DeepSeek's example of the translation removes the probability factor altogether, thus changing the possible adverse reaction into an absolute certainty. In terms of MQM, this would be classified as an Accuracy/Mistranslation error since the level of certainty is modified. This difference becomes especially important in the context of medical translation, when it is necessary to communicate the potential adverse reactions and not their inevitability. As far as Fluency is concerned, all the three examples are grammatically correct, coherent, and natural in Arabic.

Example 5: Medication Discontinuation and Treatment Adherence

This example focuses on the interpretation of instructions for stopping drugs and stresses the importance of maintaining the ban as it is.

English source text	Gemini	ChatGPT	DeepSeek
Do not discontinue the medication without consulting your physician.	لا تتوقف عن تناول الدواء دون استشارة طبيبك.	لا تتوقف عن استخدام الدواء دون استشارة الطبيب المعالج.	لا توقف الدواء قبل التحدث مع الطبيب.

All three versions successfully maintain the key instruction that forbids the cessation of treatment without the consultation of a specialist. Gemini offers a direct translation that preserves both the negative imperative and the obligation to consult a doctor. ChatGPT manages to convey the intended meaning and uses the term "الطبيب المعالج," which is a correct specification within the context. DeepSeek uses the phrase "لا توقف الدواء," which means the same thing but sounds slightly less natural compared to "لا تتوقف عن تناول الدواء." Moreover, "التحدث مع الطبيب" means talking to a doctor but sounds slightly less formal than "استشارة الطبيب." However, these minor differences do not mean that there are any actual mistakes in accuracy since the main message is preserved. In terms of MQM Fluency, Gemini and ChatGPT use more standard terminology in medicine while DeepSeek produces a version that still looks grammatically correct.

Example 6: Renal Impairment and Medication Dose Adjustment

The following example is related to the interpretation of kidney dysfunction and the need for dose adjustment depending on kidney function.

English source text	Gemini	ChatGPT	DeepSeek
Patients with renal impairment may require dose adjustment.	قد يحتاج المرضى الذين يعانون من قصور كلوي إلى تعديل الجرعة.	قد يحتاج المرضى الذين يعانون من ضعف وظائف الكلى إلى تعديل الجرعة الدوائية.	قد يحتاج مرضى الفشل الكلوي إلى تغيير الجرعة.

The translations show differences in the precision of the medical terminology used. While Gemini translates the phrase as "قصور كلوي," ChatGPT translates it as "ضعف وظائف الكلى," and both keep the idea of renal impairment. ChatGPT, in addition, specifies that this adjustment concerns the dose of medications, making it more appropriate for the context of medicine. DeepSeek uses the phrase "الفشل الكلوي," which might be interpreted as a narrower definition than the term used in the source, indicating a different medical condition. This narrowing of meaning could be considered Accuracy/Mistranslation in MQM, as it changes the population to which the recommendation will be applied. All three versions correctly preserve the modal verb may as "قد," keeping the recommendation conditional. Concerning Fluency, all three translations are grammatically correct, but the expression "تعديل الجرعة" is more common in medical literature compared to the phrase "تغيير الجرعة."

Example 7: Dyspnea and the Urgency of Medical Attention

The example below focuses on how the three systems explain how to seek urgent medical help in case of problems breathing.

English source text	Gemini	ChatGPT	DeepSeek
Seek immediate medical attention if you experience difficulty breathing.	اطلب العناية الطبية فورًا إذا واجهت صعوبة في التنفس.	اطلب المساعدة الطبية على الفور إذا كنت تعاني من صعوبة في التنفس.	اذهب إلى الطبيب عاجلًا في حال حدوث صعوبة في التنفس.

In the translation evaluation, it is clear that the terms "فورًا" used by Gemini and "على الفور" used by ChatGPT keep the urgency of the original directive intact. Moreover, both translations keep the conditional nature of the link between having difficulty breathing and requiring immediate medical attention intact. On the other hand, DeepSeek uses the more general phrase "عاجلًا الطبيب" for the urgent directive in its translation, thus failing to capture the importance of the urgent action in the original directive. This would be classified as an

Accuracy/Omission error in the MQM evaluation framework due to the lack of clinically significant temporal information. It also makes the communicative power of the directive weaker because the Arabic version of the directive does not emphasize the urgency of the directive anymore. Regarding the Fluency assessment, all the three outputs are grammatically correct and easily understandable. Nevertheless, the difference in the meaning is significant as medical directives often use temporal terms to direct patient actions.

Example 8: Elevated Liver Enzymes and Laboratory Findings

In this case, the focus is on translating laboratory results, specifically focusing on the difference between high liver enzyme levels and other information about liver function.

English source text	Gemini	ChatGPT	DeepSeek
The blood test showed elevated liver enzymes.	أظهر فحص الدم ارتفاع مستويات إنزيمات الكبد.	أظهرت نتائج تحليل الدم ارتفاعًا في مستويات إنزيمات الكبد.	أظهر تحليل الدم زيادة في وظائف الكبد.

Both Gemini and ChatGPT maintain the specific laboratory result found in the English source as "ارتفاع مستويات إنزيمات الكبد." In both cases, it shows that there were high levels of enzymes in the patient's blood and this information is the key medical information. It should be noted that ChatGPT has included the word "نتائج" that is relevant to the situation and does not change the meaning. DeepSeek's translation uses "وظائف الكبد" instead of "إنزيمات الكبد" and changes the focus from high levels of enzymes to liver functions, which results in a different concept. It should be considered an Accuracy/Mistranslation error according to MQM criteria as the translated phrase does not contain the laboratory result found in the source anymore. All three translations are grammatically correct but only DeepSeek shows how the medical term can be translated in a way that looks natural but conveys incorrect information.

Example 9: Pre-Procedural Fasting Requirements

The purpose of this example is to examine the translation of instructions concerning the need for fasting before undergoing a procedure.

English source text	Gemini	ChatGPT	DeepSeek
The patient should fast for eight hours before the procedure.	يجب على المريض الصيام لمدة ثماني ساعات قبل الإجراء.	ينبغي على المريض الصيام لمدة ثماني ساعات قبل إجراء العملية الطبية.	ينبغي للمريض عدم تناول الطعام لمدة ثماني ساعات قبل الفحص.

Gemini provides a near-literal translation of the source while preserving the fast, the eight-hour period, and the generic description of the procedure. The same is true for ChatGPT, which maintains the principal message but translates the procedure as "العملية الطبية" and may imply a narrower meaning of the original term in certain contexts. DeepSeek's translation is also informative and includes two possible semantic changes: fasting is rendered as "عدم تناول الطعام" and the procedure is translated as "الفحص." According to the MQM criteria, these semantic deviations might be classified as Accuracy/Mistranslation errors in cases when a precise fasting description is needed for medical reasons. However, as the source text does not specify that fasting includes a prohibition on drinking water, the scope of the medical procedures must not be narrowed without additional context. As for Fluency, all three translations are grammatically correct and clear in their wording.

Example 10: Severe Chest Pain and Emergency Response

In this case study, the translation of emergency directions concerning chest pains is analyzed with special focus on retaining the sense of urgency and recommended actions.

English source text	Gemini	ChatGPT	DeepSeek
If you develop severe chest pain, call emergency services immediately.	إذا شعرت بألم شديد في الصدر، اتصل بخدمات الطوارئ فورًا.	إذا أصبت بألم شديد في الصدر، اتصل بخدمات الإسعاف والطوارئ على الفور.	إذا أصبت بألم شديد في الصدر، اطلب المساعدة الطبية.

Gemini and ChatGPT both correctly preserve all the important elements of the emergency message, including the presence of chest pain, the requirement to contact emergency services, and the need for immediate action. ChatGPT's "خدمات الإسعاف والطوارئ" is a direct translation of emergency services, providing an exact description of emergency assistance without changing the message. On the other hand, DeepSeek offers an accurate but simplified translation, replacing the specific instruction to contact emergency services with "اطلب المساعدة الطبية," while omitting the word "immediately." Such omissions amount to errors of omission, and possibly mistranslation because the translated message does not specify what type of service needs to be contacted and does not emphasize the urgency of the action. Despite the fact that all translations are grammatically correct and fluent, DeepSeek shows how semantic simplification may weaken a clinical warning message.

4.2. Summary of Results

Figure 1. Summary of the Three Systems' Accuracy

Figure 1 provides a comparative view of the accuracy scores obtained by Gemini, ChatGPT, and DeepSeek in terms of their ability to translate medical information from English into Arabic. As can be seen in Figure 1, ChatGPT secured the highest score (18) compared to the other two translation services. Gemini came second with a score of 14, and DeepSeek obtained the lowest score of 10. In this regard, ChatGPT scored higher than its competitors in the aspect of accuracy depicted in Figure 1. Consequently, ChatGPT was more successful than the other two in preserving the meaning and the content of the medical information. While Gemini proved to have comparatively moderate accuracy, DeepSeek lagged behind in terms of translation accuracy. However, such differences depend on the method of evaluation used, and the provided figure cannot be interpreted as an indication of statistical significance of the differences observed.

5. Discussion

This paper explores the efficiency of Gemini, ChatGPT, and DeepSeek in terms of their ability to translate medical data from English to Arabic using the Multidimensional Quality Metrics (MQM). The comparison reveals the ability of these generative AI models to provide fluently written Arabic translations of medical texts. At the same time, there are differences between them in preserving the medical terminology, clinical probabilities, diagnoses, and emergency information. Thus, it is obvious that linguistic fluency does not guarantee semantic accuracy especially in medical translation.

This study reveals one of the strengths of the three models in their ability to generate grammatically correct Arabic translations of relatively simple medical sentences, such as doses of medications and their frequency of use, without losing the required information. Similarly, Gemini and ChatGPT reveal strong performance in the preservation of medical terminology in different examples provided, while some DeepSeek examples use a general or vague equivalent of the same terms. This supports the claims made by Costa-jussà et al. (2022), as the researchers have stressed the development of multilingual machine translation, but it also highlights the problems mentioned by Vieira et al. (2021).

Another interesting finding refers to the way modalities, negations, and emergency instructions were translated. As an example, cases of possible medication side-effects, previous experience with diabetes, difficulty breathing, and severe chest pain show that failure to translate certain phrases, such as may or immediately, can change the meaning of the medical information being given. This finding is corroborated by the WHO (2024) recommendations that refer to the reliability and safety of healthcare-related AI applications, requiring human supervision. How far users can trust such output has also been examined in reception research on AI translation (Haider, 2026b). Furthermore, the findings of this study are supported by other studies examining machine translation of emergency discharge instructions carried out by Carreras Tartak et al. (2025).

It should also be noted that the research reveals the significance of distinguishing between Accuracy and Fluency in evaluating AI translations. There are several examples of translations that are grammatically correct but might contain significant semantic errors. The results obtained in the course of the study confirm the claims made by Freitag et al. (2021) regarding the relevance of using expert analysis of errors for evaluation of machine translation. The application of MQM in the current study provides a structured approach to the identification of mistranslation, omission, and terminological errors.

Furthermore, the paper also helps build upon the literature on voice-translation technology with the

help of AI, through its focus on medical speech translation, which combines speech recognition and translation techniques. The improvements in multilingual speech recognition and speech translation technology made possible by Radford et al. (2023) and Barrault et al. (2023), respectively, showcase the possibilities of these technologies. Yet, due to the possibility of mistakes in the recognition phase affecting the following translation process, it is important to distinguish between these two.

In summary, the major strengths of the study lie in its comparative analysis of three AI generators, the implementation of the MQM model, and the incorporation of linguistic analysis alongside clinical translation aspects. The analysis reveals the possibilities for employing the tools of Gemini, ChatGPT, and DeepSeek in English–Arabic medical translation but stresses the need for human revision, especially in terms of terminology, prescriptions, and emergencies. Further research must apply medical audio databases and independently performed clinician-based evaluations of translation errors to verify the trustworthiness of the analyzed systems.

6. Conclusion

The current research explored the efficacy of three different generative AI systems, namely, Gemini, ChatGPT, and DeepSeek, to translate medical information from English to Arabic through the MQM approach. The research was based on the two key parameters of translation quality, which are accuracy and fluency, paying special attention to the medical terms, diagnostic information, medicinal instructions, and the terminology related to emergency situations. It can be seen from the comparative examples above that all the systems have the potential to provide grammatically acceptable and understandable Arabic translations. Nevertheless, there were certain differences between the three systems in terms of the precision of terminology usage, the preservation of meaning and clinical information, which show that translation fluency alone cannot be used as an indicator of medical translation quality.

The comparative accuracy scores illustrated in Figure 1 show that the highest score (18) was received by ChatGPT, whereas the second and third positions were occupied by Gemini (14) and DeepSeek (10), respectively. As higher scores indicate greater accuracy, ChatGPT proved to be the most accurate system, Gemini displayed moderate accuracy, and DeepSeek scored the lowest. These scores reflect the MQM-based evaluation applied in this study and should be read alongside the qualitative analysis. The key point of this study is the comparative analysis of three AI systems in English–Arabic medical translation and the use of MQM to differentiate between semantic accuracy and linguistic fluency.

Although this study shows promising opportunities for these systems in the future, there are some issues related to the scope of the corpus used in the analysis, medical topics covered, and the generalization of results. Moreover, the qualitative analysis focused on ten representative examples, and future studies could extend this analysis to larger samples. Future studies should focus on working with bigger data sets covering other medical specialties, speakers' accents, and different recording conditions, with independent expert evaluation and a measure of inter-rater reliability included. Additionally, more work needs to be done in order to determine whether errors are made during the speech recognition process or in the process of translation. In general, generative AI systems present good opportunities for English-Arabic medical communication.

Funding Statement
This research received grant no. (330) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

References

Almahasees, R., & Almahasees, Z. (2025). Breaking language barriers in physiotherapy: Ensuring consistent English-to-Arabic translations for better patient care. *Research Journal in Advanced Humanities*, *6*(2), 1-12.

Almahasees, Z. (2023). A Synchronic Assessment of Google Translate, Microsoft Translator and Reverso in English<> Arabic Translation. In *3rd International Conference on Language and Education* (pp. 1-6). Cihan University-Erbil.

Almahasees, Z., & Husienat, I. (2024). A comparative analysis of terminological inconsistency in scientific translation from English into Arabic across different medical fields. *Training, Language and Culture*, *8*(3), 25-40.

Almahasees, Z., & Husienat, I. (2026). AI-Powered Voice Translation in Medical Settings: An Evaluation of English–Arabic Translation Accuracy. *International Journal of Linguistics, Literature and Translation*, *9*(9), 203-214.

Barrault, L., Chung, Y.-A., Meglioli, M. C., Dale, D., Dong, N., Duquenne, P.-A., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., Klaiber, C., Li, P., Licht, D., Maillard, J., Rakotoarison, A., Sadagopan, K. R., Wenzek, G., Ye, E., ... Wang, S. (2023). *SeamlessM4T: Massively multilingual & multimodal machine translation [Preprint]*. arXiv. https://doi.org/10.48550/arXiv.2308.11596

Boulesnam, I., & Boucetti, R. (2025). Arabic Language Characteristics that Make its Automatic Processing Challenging. *International Arab Journal of Information Technology (IAJIT)*, *22*(4), 814.

Carreras Tartak, J. A., Haimovich, A. D., Thomas, C. E., Bivens Moore, E., Slebonick, A., Haque, N., & Thomas, S. H. (2025). Evaluating the accuracy of Google Translate in emergency department discharge instructions. *The Journal of Emergency Medicine*, 79, 132–142. https://doi.org/10.1016/j.jemermed.2025.04.024

Delfani, J., Orasan, C., Saadany, H., Temizoz, O., Taylor-Stilgoe, E., Kanojia, D., ... & Schouten, B. (2024). Google translate error analysis for mental healthcare information: Evaluating accuracy, comprehensibility, and implications for multilingual healthcare communication. *arXiv preprint arXiv:2402.04023*.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460–1474. https://doi.org/10.1162/tacl_a_00437

Haider, A. S. (2019a). The representation of Al-Megrahi's release in Arabic and English newspapers in 2009 and 2010: A corpus-assisted discourse study. *Dirasat: Human and Social Sciences, 46*(1, Suppl. 2), 297–317. https://doi.org/10.35516/0103-046-986-019

Haider, A. S. (2019b). Syrian–Lebanese relations: A corpus-based critical discourse analysis of Bashar Al-Assad's speeches and interviews. *Dirasat: Human and Social Sciences, 46*(4), 551–570. https://doi.org/10.35516/0103-046-004-040

Haider, A. S. (2019c). Using corpus linguistic techniques in (critical) discourse studies reduces but does not remove bias: Evidence from an Arabic corpus about refugees. *Poznan Studies in Contemporary Linguistics, 55*(1), 89–133. https://doi.org/10.1515/psicl-2019-0004

Haider, A. S. (2026a). AI as a language tutor: Jordanian university students' perceptions of AI-generated feedback on their English. *Research Journal in Advanced Humanities, 7*(4), 1–17.

Haider, A. S. (2026b). Trust in AI translation: Reception and adoption among Jordanian business professionals. *Research Journal in Advanced Humanities, 7*(4), 1–15.

Haider, A. S., Abu Akar, R., Saed, H., Abu Tair, S., & Ismail, H. H. (2025). AI-powered Arabic translation of business and finance acronyms and abbreviations: Evidence from Gemini, ChatGPT, and Google Translate. In *Innovation, technologies, and business management (ICTIM)* (pp. 519–528). Springer. https://doi.org/10.1007/978-3-031-84889-6_62

Haider, A. S., & AlKhatib, R. (2024). Subtitling English legal acronyms into Arabic: Human vs machine. *Kutafin Law Review, 11*(4), 810–842. https://doi.org/10.17803/2713-0533.2024.4.30.810-842

Haider, A. S., Saed, H., Albarakati, M., Abu Tair, S., & Jarrah, S. (2025). Phonetic and orthographic transliteration of borrowed words in Arabic: A translation-oriented study. *WORD, 71*(1), 1–19. https://doi.org/10.1080/00437956.2025.2455229

Hartung, T., & Kleinstreuer, N. (2025). Challenges and opportunities for validation of AI-based new approach

methods. *Altex*, *42*(1), 3-21.

Ismayilli, T. (2024). Navigating complexities in medical text translation: Challenges, strategies, and solutions. *Acta Globalis Humanitatis Et Linguarum*, *1*(2), 170-176.

Karakus, I. S., Strechen, I., Gupta, A., Nalaie, K., Chen, C. L., Hassett, L. C., & Barwise, A. K. (2025). Bridging language gaps in healthcare: a systematic review of the practical implementation of neural machine translation technologies in clinical settings. *Journal of the American Medical Informatics Association*, *32*(11), 1756-1766.

Karanikolas, N., Manga, E., Samaridi, N., Tousidou, E., & Vassilakopoulos, M. (2023). Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics* (pp. 278-290).

Khujakulov, R. (2026). Semantic transformations of Arabic-origin medical lexemes in "The Canon of Medicine": A case study of Uzbek folk medicine and English equivalents. *Asian journal of medical sciences and clinical research*, *1*(1), 42-48.

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, *16*(4), 1-23.

Li, D. (2026). Post-editing strategy optimization and performance evaluation based on DQF-MQM error analysis. *Discover Applied Sciences*, *8*(4), 357.

Lommel, A., Gladkoff, S., Melby, A. K., Wright, S. E., Strandvik, I., Gasova, K., ... & Nenadic, G. (2024, September). The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)* (pp. 75-94).

Mahdi, A. I., & Mohammed, H. G. (2025). Accuracy of medical terminologies translated from English into Arabic: The case of audiovisual shows. *International Journal of Arabic-English Studies*, 147-172.

Mejaški, A. (2026). *A Comparative Analysis of AI Translation Errors in Medical and Legal Texts* (Master's thesis, Sveučilište u Rijeci, Sveučilište u Rijeci, Filozofski fakultet, Katedra za translatologiju).

Mingote, V., Gimeno, P., Vicente, L., Khurana, S., Laurent, A., & Duret, J. (2023). Direct text to speech translation system using acoustic units. *IEEE Signal Processing Letters*, *30*, 1262-1266.

Muhanna, A. A. I. (2025). Assessing machine translation quality on uterine cancer content: MQM-based comparative study. *CDELT Occasional Papers in the Development of English Education*, *89*(1), 421-458.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., ... Wang, J. (2022). *No language left behind: Scaling human-centered machine translation [Preprint]*. arXiv. https://doi.org/10.48550/arXiv.2207.04672

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 28492–28518). *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v202/radford23a.html

Sapawi, M. S. M., & Yusoff, N. M. R. N. (2026). Arabic Language and Medical Terminology in Education: A Systematic Review. *IIUM Medical Journal Malaysia*, *25*(01).

Trigka, M., & Dritsas, E. (2025). The evolution of generative AI: Trends and applications. *IEEE access*, *13*, 98504-98529.

Vieira, L. N., O'Hagan, M., & O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: A critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11), 1515–1532. https://doi.org/10.1080/1369118X.2020.1776370

World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. https://www.who.int/publications/i/item/9789240084759

Yelahina, N. I., & Fedchyshyn, N. O. (2025). *The role of applied linguistics in enhancing medical translation accuracy: a corpus-based approach*. https://dspace.luguniv.edu.ua/xmlui/handle/123456789/13249