Research Journal of English Language and Literature (RJELAL)
A Peer Reviewed (Refereed) International Journal
Impact Factor 6.8992 (ICI) http://www.rjelal.com;
Email:editorrjelal@gmail.com; ISSN:2395-2636 (P); 2321-3108(O)

Vol.14.Issue 3. 2026 (July-Sept.)

RESEARCH ARTICLE

INTERNATIONAL STANDARD SERIAL NUMBER INDIA
2395-2636 (Print):2321-3108 (online)

Semantic Accuracy and Cultural Equivalence in AI Translation of Jordanian Proverbs: A Comparative Evaluation of DeepSeek, Claude, and Grok

Laith Hmoud Abdel Hafiz Allaymoun[1] and Dr. Rakesh Cherucodu[2]
[1]Department of Linguistics, Central Institute of Indian Languages, University of Mysore, Mysuru, Karnataka, India
Email: mactop1327@gmail.com
[2]Principal, Southern Regional Language Centre, Central Institute of Indian Languages, University of Mysore, Mysuru, Karnataka, India

DOI: 10.33329/rjelal.14.3.345

Article info
Article Received: 10/08/2026
Article Accepted: 06/09/2026
Published online: 11/09/2026

Abstract

This study examines the semantic accuracy and cultural equivalence of English translations of Jordanian proverbs generated by three AI models: DeepSeek, Claude, and Grok. The study analyses 190 Jordanian proverbs, producing a corpus of 570 AI-generated translations. Drawing on Nida's theory of equivalence, semantic accuracy and cultural equivalence were evaluated as two distinct dimensions using five-point evaluation rubrics. A quantitative and qualitative approach was employed, combining descriptive analysis, the Friedman test, pairwise Wilcoxon signed-rank tests, and qualitative examination of representative translations. The findings showed generally high semantic accuracy across the three models, with Claude producing the highest proportion of Highly Accurate translations (78.95%), followed by DeepSeek (66.32%) and Grok (64.21%). Significant differences were found among the models ($p < .001$), with Claude significantly outperforming both DeepSeek and Grok, while the latter two did not differ significantly. Greater variation emerged in cultural equivalence. Claude achieved 62.11% Highly Culturally Equivalent translations, compared with 38.42% for DeepSeek and 34.74% for Grok, and again significantly outperformed both models ($p < .001$). The qualitative findings revealed that literal translation did not necessarily reduce semantic accuracy when the intended meaning remained recoverable. However, preserving source imagery did not necessarily ensure cultural equivalence. Cultural and semantic losses occurred particularly when the models misinterpreted Jordanian dialectal expressions, figurative meanings, or culturally embedded associations. The findings demonstrate that semantic accuracy and cultural equivalence are related but distinct dimensions of AI translation quality: a

translation may successfully communicate the general meaning of a proverb while failing to convey its cultural significance. The study highlights the need to evaluate AI translation beyond surface linguistic accuracy and to improve models' ability to interpret dialectal, figurative, and culturally embedded proverbial language.

Keywords: Artificial Intelligence; Machine Translation; Jordanian Proverbs; Semantic Accuracy; Cultural Equivalence; DeepSeek; Claude; Grok.

INTRODUCTION

In recent years, artificial intelligence (AI) has gained significant importance in the translation industry, especially as large language models have emerged to handle intricate language and context. Recent AI translation models are able to produce fluent translations, understand context, and adhere to complex translation directives, which is something that traditional machine translation systems cannot do. With these advancements, AI translation has evolved from simple informational text to more intricate language, such as figurative, cultural, and context-specific expressions (Alhumsi et al., 2025; Naveen & Trojovsky, 2024). However, it should be noted that while language proficiency is important, AI models are not always capable of capturing all aspects of meaning, especially in the context of culturally bound expressions like proverbs.

Proverbs pose a more difficult type of translation, in which the meaning is frequently found in the interplay of the words, rather than in the literal translation of each word. They are short statements that contain shared knowledge, beliefs, and culturally shared values and experiences (Mieder, 2004). The interpretation of their meaning is therefore a task that requires knowledge of the language and the culture of the "language." A translation can be a faithful rendition of the meaning of a proverb without affecting its overall message, but may lose its cultural undertones, figurative language, or implications. Therefore, evaluating proverb translation demands care not only for the attainment of the meaning of the source language proverb but also for the ability to transfer the cultural implication of the source language proverb.

This difference is particularly noticeable in the translation of Arabic to English. Occasionally, culturally specific images, religious references, social relationships, and traditional beliefs exist in Arabic proverbs that do not necessarily have equivalents in the English language. Jordanian proverbs are especially relevant, as many are expressed in colloquial speech and are indicative of social practices, values, historical experiences, and culturally bound meanings (Al-Azzam, 2018; Alarsan & Khan, 2025). Al Rousan and Shatnawi (2023) reported that despite the fact that the translation of Jordanian Arabic proverbs into English is straightforward, it still faces a series of linguistic, translation and comprehension problems such as lexical problems, literal rendering, semantic distortion and problems with figurative and cultural meanings. Likewise, Al Kayed et al. (2023) noted that cultural differences and the connotative meaning of Jordanian proverbs were the great challenges for translators.

These difficulties raise two closely related issues: semantic accuracy and cultural equivalence. In semantic correctness, the correctness of the translated meaning of the source proverb is discussed, without significant loss, distortion or misunderstanding. Cultural equivalence, on the other hand, refers to how well the translated text preserves or communicates culturally relevant meaning, associations, imagery and implications. The

difference is crucial because high semantic accuracy does not guarantee high cultural equivalence. A translation can convey the general meaning of a proverb without capturing the cultural associations through which that meaning can be felt.

This relationship could be explained by Nida's (1964) theory of equivalence, as it deals with two aspects of the theory: first, formal correspondence and second, orientation towards the production of an equivalent response in the target language. The most important thing for Nida was to stress that translation is not merely reproduction of linguistic form, but should also take into account the functioning of meaning for the target audience. This view is especially applicable to the translation of proverbs, as the words or figurative pattern of the original expression might not have the same effect in the target language. Thus, Nida's theory is a suitable theoretical framework for analyzing the extent to which AI translations maintain the intended meaning and also convey culturally embedded meaning.

Recent studies have shown the capability and drawbacks of AI in this realm. Large language models like ChatGPT have often outperformed traditional machine translation systems in generating more natural and contextually appropriate translations, especially with figurative expressions (Abdelhalim et al., 2025; Aldawsari, 2024). But AI translations still struggle with culturally specific expressions, metaphors, idioms and culturally embedded meanings. Other studies, such as Benyahia (2025), have revealed that ChatGPT could provide fluent translations but was generally more literal with the culturally specific and opaque expressions in Arabic. Al-Kaabi et al. (2024) also stated that AI tools could generate fluent translations, but they might miss out on cultural imagery, metaphorical meaning, and stylistic elements. These are indications that assessing AI translation quality based on fluency is not enough.

The difficulty of such a challenge can be exacerbated by the inclusion of Arabic dialects. There is a significant difference between Standard Arabic and other local forms of Arabic, and the ability to perform Standard Arabic cannot be generalized to local (dialectal) expressions (Ali, 2016). Another study by Kadaoui et al. (2023) showed that the model performance differs between Arabic varieties, underlining the ongoing challenges of dialectal and low-resource language processing. Jordanian proverbs incorporate this dialectal dimension, along with culturally embedded proverbial meaning, and serve as a fruitful and testing example of the ability of contemporary AI models to go beyond surface-level linguistic transfer.

Although Arabic–English translation has received significant attention in the literature on AI-assisted translation, few studies have focused on the comparative analysis of across-the-board Arabic–English translation using recently developed AI models, especially with respect to the complete assessment of both the semantic and cultural aspects. Furthermore, models can vary significantly if they are used to process the same expression, which requires a direct comparison of the models (Obeidat et al., 2024). To fill this gap, this study aims to examine the semantic accuracy and cultural equivalence of 570 English translations of 190 Jordanian proverbs generated by DeepSeek, Claude, and Grok. Based on Nida's theory of equivalence, it analyzes the degree of semantic accuracy and cultural equivalence of the three models and finds out whether there are statistically significant differences between the performances of the three models. The study offers evidence of the linguistic and cultural skills of the current AI systems in their ability to handle culturally embedded Jordanian proverbial language.

Nature and Translation of Jordanian Proverbs

In the assessment of proverb translation, semantic accuracy and cultural equivalence are

two aspects that are closely related but different. A translation can be right in meaning while not reflecting the cultural context, metaphoric references or social history of the source proverb. This difference is especially relevant when translating Jordanian proverbs to English, as the meaning of the proverbs often goes beyond the meaning of the individual words and relies on the shared cultural background and context of the target culture (Alarsan & Khan, 2025; Maalej, 2009). Thus, a translation of proverbs by AI algorithms must be considered in terms of its meaning conveyed as well as its fidelity in representing the cultural meaning of the original.

Semantic accuracy is the degree of fidelity of the translation in retaining the meaning of the source expression without semantic loss, distortion and misunderstanding. In the case of proverb translation, this means that one must be able to see that the meaning could be figurative instead of literal. If a proverb's literal meaning is not understood, then the translation of its lexical content can be correct, but the meaning of the proverb will not be captured. For instance, Benyahia (2025) noted that while ChatGPT often translated Arabic idioms into a fluent and grammatical translation, sometimes the literal translation of the idioms rendered an opaque and culture-specific expression into a different one with a different meaning. Likewise, Al Rousan and Shatnawi (2023) pointed out that the distortion of meaning, inaccurate lexical selection, literal translation, and problems with understanding figurative meanings are among the general problems faced in translation from Jordanian Arabic to English.

Therefore, two critical tasks should be distinguished: semantic accuracy and cultural equivalence. The translated, natural English phrase may not be a true reflection of the source phrase. AI translations were found to be fluent yet also contained errors (mistranslations or omissions) that impacted their semantic appropriateness (Kadhim, 2024). Cahyaningrum (2024) also noted that increased fluency does not necessarily lead to increased semantic fidelity, as AI may focus on naturalness rather than on close fidelity to meaning. These results validate the measure of semantic accuracy as a separate dimension from grammatically natural output, which is not necessarily an indicator of success in translation.

Cultural equivalence, on the other hand, is related to the degree of preservation or appropriate communication of the cultural meaning, associations, imagery and implications of the source expression. Proverbs rely particularly on these elements, as they are linked to the values, experiences and common knowledge of the community. According to Hamdi et al. (2023), an Arabic proverb can be similar in form and content to an English proverb, but have a different religious, cultural or pragmatic implication. Similarly, Al-Azzam (2018) demonstrated that Jordanian proverbs are not only socially, historically and culturally specific but also cannot be translated literally without capturing the meaning.

The difference between the accuracy of meaning and cultural equivalence is also apparent in studies of AI translation. Al-Kaabi et al. (2024) stated that AI systems potentially could produce fluent and natural translations without retaining cultural images, metaphors and connotations. Similarly, Mohsen (2024) discovered that sometimes, ChatGPT retained the surface meaning but lacked the depth and nuance of Arabic cultural elements. Therefore, if culturally meaningful images or associations are either made less meaningful or omitted, the resulting translation of a proverb may be able to capture its main message, but it will not be culturally equivalent.

The connections among these dimensions are especially important in Nida's (1964) theory of equivalence. By distinguishing formal and dynamic equivalence, Nida takes translation evaluation a step further from linguistic correspondence to the communicative effect of the translated expression. It is not enough to

produce the words or structures of the source language for culturally embedded language; it must be communicated in a form that will be suitable for target-language readers. This is similar to Alharbi (2023), who concluded that dynamic equivalence outperforms literal translation in retaining the intended meaning, emotional impact and pragmatic use of Dialectal Arabic proverbs.

Accordingly, the present study treats semantic accuracy and cultural equivalence as separate but complementary evaluation dimensions. In order to evaluate the semantic accuracy of DeepSeek, Claude, and Grok, we assess their ability to capture both the meaning behind the Jordanian proverbs and the culturally embedded meanings and associations that Jordanian people would make out of them. Both dimensions are examined and offer a wider context for assessing whether the meaning of Jordanian proverbs, as they are embedded in the culture, can be conveyed to English by contemporary AI models beyond surface linguistic transfer.

Since semantic accuracy and cultural parallelism are two distinct but complementary aspects of translation, discrepancies can arise in the performance of each AI model with respect to each of these two aspects. There have been a number of previous studies that demonstrated that the performance of AI translation models differs depending on the model and evaluation method, especially when dealing with figurative and culturally embedded expressions (Ahmed et al., 2025; Al-Jarf, 2025a; Obeidat et al., 2024). In this regard, the present study aims to assess the degree of semantic accuracy and cultural equivalence that DeepSeek, Claude and Grok could achieve, as well as statistically determine whether they differ significantly. Based on the above, the following null hypotheses are developed:

H01: There is no statistically significant difference in semantic accuracy among DeepSeek, Claude, and Grok.

H02: There is no statistically significant difference in cultural equivalence among DeepSeek, Claude, and Grok.

REVIEW OF THE LITERATURE

The translation of proverbs is not just a word-for-word translation but is closely related to figurative language, shared cultural knowledge and the pragmatic context. This problem is especially applicable to the case of Arabic-English translation. Maalej (2009) states that Arabic proverbs carry cultural knowledge, social values and shared worldviews and, therefore, the surface level of the linguistic transfer may not convey the cultural meaning. In the same manner, Hamdi et al. (2023) identified that although Arabic and English proverbs may have similar themes or structures, they do not necessarily have the same cultural, religious and pragmatic meaning. Thus, the equivalence of expressions does not necessarily follow from similarity of language.

The challenge is more specific in the translation of Jordanian proverbs. Al-Azzam (2018) found that Jordanian proverbs are closely connected with their social and historical background, as well as their colloquial nature, adding to the challenges of translation. Literal translation can then not necessarily maintain their semantic and cultural meanings. More recently, Alarsan and Khan (2025) highlighted that Jordanian proverbs and idioms have culturally embedded meanings related to Jordanian values, beliefs, history and social contexts. Empirical evidence also shows that it can be challenging to preserve these meanings. When translating Jordanian Arabic proverbs into English, Al Rousan and Shatnawi (2023) found that there were lexical errors, meaning distortion, literal translation, and difficulties in understanding figurative and culture-specific meanings. Al Kayed et al. (2023) also revealed that cultural differences and the lack of awareness of the connotative meaning were the main reasons for the failure in the translation of Jordanian proverbs.

The difficulties these pose create a crucial aspect to assess the quality of AI translations – semantic accuracy. Semantic accuracy refers to the absence of meaningful loss, distortion or misinterpretation from the source expression's intended meaning. New research shows that fluent AI output does not always lead to semantic accuracy. Benyahia (2025) discovered that ChatGPT often produced grammatically correct and fluent translations of Arabic idioms, although sometimes expressions translated into Arabic had distorted meanings. Kadhim (2024) also noted that while AI-generated translations might be fluent, they could also contain mistakes or leave out parts of the message, potentially skewing the intended meaning. Semantic accuracy, fidelity, adequacy, and fluency were among the major criteria used to assess the performance of Arabic–English machine translation, as found in Saeed's (2025) more narrowly focused systematic review.

Cultural equivalence presents an additional and potentially more complex challenge. A translation may communicate the basic semantic message while weakening cultural imagery, associations, or pragmatic significance. Al-Kaabi et al. (2024) showed that AI systems could provide natural translations but with a lack of cultural imagery, metaphorical meaning, and cultural connotations. Similarly, Khoshafah (2024) found that ChatGPT generated coherent Arabic-English literary translations but struggled to maintain the cultural nuances, figurative expressions, and intertextual references in the translation. These results show that semantic success and cultural success cannot be considered as one and the same measure of the quality of translation.

There are also significant differences between AI models, according to research. Similarly, Al-Jarf (2025a) revealed some discrepancies between Microsoft Copilot, DeepSeek, and Google Translate when analyzing metaphorical expressions, with DeepSeek often producing literal translations and struggling with culturally specific meanings. Additionally, Obeidat et al. (2024) showed a significant variation in the performance of Google Translate, ChatGPT, and Gemini in processing idiomatic expressions. Recently, Paphonphat (2026) found significant variations in the accuracy of the translated content between ChatGPT, Gemini, DeepSeek, and Claude when translating idiomatic content. These results imply that the performance of a single AI model does not necessarily carry over to another model.

There is an important caveat in the above research, however. Previous research has focused on the translation of Jordanian proverbs by human translators, translation of Arabic proverbs by human translators, or the evaluation of figurative and culturally embedded language in different contexts. The reviewed literature, however, does not allow a direct comparative evaluation of DeepSeek, Claude and Grok in terms of translating Jordanian proverbs into English from two different evaluation dimensions (semantic accuracy and cultural equivalence). This is crucial as Jordanian proverbs are spoken in dialectal language and carry culturally embedded and figurative meanings which might vary significantly from one AI model to another. This study then assesses 570 translations, based on 190 Jordanian proverbs, to see how well the three models can capture the intended meaning and cultural significance.

Based on this gap, the study addresses the following research question: To what extent do the English translations of Jordanian proverbs generated by DeepSeek, Claude, and Grok achieve semantic accuracy and cultural equivalence?

Nida's Theory of Equivalence and Evaluation Framework

Nida's Theory of Equivalence provides the theoretical foundation for understanding the transfer of meaning between source and target languages. Based on this perspective, the

present study uses two evaluation rubrics to systematically assess semantic accuracy and cultural equivalence in the AI-generated translations.

Nida's Theory of Equivalence

Nida's Theory of Equivalence provides the theoretical basis for examining the semantic accuracy and cultural equivalence of the English translations generated by DeepSeek, Claude, and Grok. It is especially applicable to the translation of proverbs as it transcends the correspondence between language forms and places the emphasis on meaning communication and the understanding of the target-language reader. For the purposes of Jordanian proverbs, it is significant that the meanings intended may rely on figurative language, cultural associations and common background knowledge of society, which may not always be conveyed through literal translation of the words.

Nida (1964) separated Equivalence into two classes: Formal and Dynamic. Formal equivalence is source-oriented, meaning that it aims at preserving the form and content of the source text in the target text. Lexical items, grammatical forms and images of metaphor may, therefore, be retained in a formally equivalent translation. This orientation can help the reader in the English translation to see the image of the original proverb in Jordan. But the purposeful reproduction of the linguistic form does not guarantee that the purposeful meaning and cultural content of the original language will be naturally understood in the target language.

Dynamic equivalence, on the other hand, emphasizes that the source message be communicated naturally and appropriately to the receptor in the target language. Nida and Taber (2003) highlight the reproduction of the nearest equivalent to the source-language message, meaning takes priority over style. This orientation allows for variations in wording, structure, and/or imagery when direct reproduction would not communicate the intended message. For proverbs, an English expression can thus be structurally different from the source proverb in Jordan but have a similar meaning and function. Formal and dynamic equivalence should consequently be viewed as different orientations towards translation rather than simple measures of correct and incorrect translation.

Nida's approach later developed towards the concept of functional equivalence. de Waard and Nida (1986) adopted the term to focus more on the communicative roles served by the translated language. Functional equivalence is based upon a set of principles about the receptor, rather than being a different kind of equivalence. It's all about whether the translation conveys the message's meaning and action, or merely its language structure. This is particularly valuable for culturally bounded expressions as their communicative value may be culturally contingent and rely on conventional connections that are not necessarily available from word-for-word correlation.

Nida's theory offers an interpretive framework for the two major dimensions of evaluation in the present study: semantic and cultural. Semantic accuracy relates to how well the meaning of a Jordanian proverb is conveyed in English with little distortion, omitting, adding, or misinterpreting significant items. Cultural equivalence does not only involve the meaning of the words, but also whether the culturally embedded meaning, association, imagery, and communicative significance are sufficiently represented. This distinction is important because a translation can convey the overall meaning of a proverb but at the same time diminish an important cultural connection. On the other hand, it could retain the original cultural image but not effectively convey the meaning of the image to the English reader.

In this study, Nida's theory is not applied as a numerical "scoring scale." Instead, it offers a

theoretical lens on the scores for cultural equivalence and semantic accuracy. It allows for analysis to identify whether DeepSeek, Claude, and Grok are just holding onto aspects of the Jordanian source expression or if they are sharing the intended meaning and culturally embedded meaning of the expression well in English. In this manner, Nida's theory is an appropriate foundation for comparison of the capability of the three AI models to go beyond linguistic correspondence to meaning and culturally appropriate proverb translation.

Semantic Accuracy Evaluation Rubric

The semantic accuracy was evaluated with a rubric developed for the present study, which is based on a five-point scheme. The rubric criteria assess how well each AI-generated translation captures the meaning of the Jordanian proverb, including the literal and figurative meanings, while avoiding significant semantic loss, distortion, or misinterpretation.

Table 1: Semantic Accuracy Evaluation Rubric

Score	Level	Operational Criterion
5	**Highly Accurate**	Fully conveys the intended meaning with no meaningful semantic loss or distortion.
4	**Accurate**	Conveys the central meaning accurately with only minor semantic differences.
3	**Moderately Accurate**	Conveys the general meaning, but some semantic or figurative elements are weakened or lost.
2	**Slightly Accurate**	Preserves only a limited part of the intended meaning, with substantial semantic distortion.
1	**Inaccurate**	Fails to convey or substantially misrepresents the intended meaning.

Table 1 shows the five-point rubric used to assess the semantic accuracy of the AI-generated translations. The scores range from 1: Inaccurate to 5: Highly Accurate. The assessment is based on how well the intended meaning of each Jordanian proverb is conveyed in English.

Cultural Equivalence Evaluation Rubric

Cultural equivalence was assessed independently with a rubric created for this study that had 5 points. It assesses the extent to which each translation captures or effectively communicates the cultural inclusion of meaning, association, imagery and implications of the Jordanian proverb.

Table 2: Cultural Equivalence Evaluation Rubric

Score	Level	Operational Criterion
5	**Highly Culturally Equivalent**	Fully preserves or appropriately conveys the cultural meaning and relevant associations.
4	**Culturally Equivalent**	Preserves the central cultural meaning with only minor differences.
3	**Moderately Culturally Equivalent**	Conveys the general cultural meaning, but some imagery or associations are weakened or lost.

2	**Slightly Culturally Equivalent**	Preserves only a limited part of the cultural meaning, with substantial cultural loss or modification.
1	**Culturally non-equivalent**	Fails to convey or substantially misrepresents the culturally intended meaning.

Table 2 shows the five-point rubric that was used to assess cultural equivalence from the lowest (Culturally Non-Equivalent) to the highest (Highly Culturally Equivalent). The evaluation centers on how well the cultural significance, connotations, and implications of each of the Jordanian proverbs are captured in English.

Nida's theory thus serves as a theoretical framework for the analysis of equivalence, and the two sets of evaluation rubrics describe equivalence in the 570 translations generated by AI in a systematic way by evaluating them for semantic accuracy and cultural equivalence.

MATERIALS AND METHODS

This section describes the methodological procedures used to evaluate and compare the semantic accuracy and cultural equivalence of Jordanian proverb translations generated by DeepSeek, Claude, and Grok. It covers the research design, corpus, translation procedure, evaluation process, expert review, and data analysis.

Research Design and Corpus

The descriptive-comparative approach was applied in this study using both quantitative and qualitative methods to assess and compare the semantic accuracy and cultural equivalence of the translated Jordanian proverbs produced by DeepSeek, Claude, and Grok. This was due to the specific characteristics of the corpus (purposive sampling): it was developed specifically to include proverbs recognised within the Jordanian community and used in it.

The source material of the corpus was Jordanian Arabic proverbs, a total of 190, which were collected mainly from oral and community sources. Traditional Jordanian proverbs were collected from a group of older Jordanian men and women aged 40+, who knew their meaning and how they are used. The researcher's knowledge of the Jordanian proverbs helped to develop the first draft of the corpus, while the inclusion of the proverbs in Jordanian contexts was also verified by people who have native knowledge of Jordanian proverbial expressions. The corpus included various facets of daily life, social interactions, cultural experience, and traditional knowledge. All 190 proverbs have been kept for analysis.

Translation Procedure

The models tested were DeepSeek-V4-Pro, Grok 4.5, and Claude Opus 5. The translation data were generated between 30th July 2026 and 25th July 2026. All 190 proverbs were presented one by one to each of the 190 models, under the same prompting conditions, resulting in 190 translations per model and a total of 570 English translations. For each of the three models, the same standardised prompt was used:

"Translate the following Jordanian Arabic proverb into English. Preserve the original meaning as accurately as possible. Preserve the cultural meaning whenever possible. Produce only one English translation. Do not provide explanations, notes, or alternative translations. Output only the translated proverb. Jordanian proverb: [proverb]."

The models were not given an explanation of the meaning or cultural context of the proverbs. The models were not asked to change their first production, and the

translations were not post-edited; they were written as they were generated.

Coding, Expert Review, and Data Analysis

The study developed five-point evaluation rubrics for each evaluation, and all 570 AI-generated translations were evaluated separately in terms of semantic accuracy and cultural equivalence. Semantic accuracy was judged by determining to what degree each translation captured the meaning of the source proverb without losing any meaning, distorting, or misinterpreting the meaning. The cultural equivalence of the Jordanian proverb was evaluated on how much the translation preserved or matched the cultural meaning, association, imagery and implications of the Jordanian proverb.

The researcher first assessed the translations, and the whole data set was subsequently assessed by three evaluators specializing in translation, English and linguistics. The evaluators reviewed the scores assigned, and agreed-upon scores were assigned by discussion of any disagreement that arose.

Representative translations were purposively selected from the full translation set for the qualitative analysis to highlight salient patterns of meaning loss, modification, and preservation and cultural differences between the three models. All the statistical results were calculated from a complete data set of 570 translations, and the qualitative examples were employed to explain and support the quantitative results.

Data Analysis

The semantic accuracy and cultural equivalence scores of DeepSeek, Claude, and Grok were summarized as descriptive statistics such as frequencies, percentages, means, and standard deviations. To find out if there were statistically significant differences between the three models for each evaluation dimension, the Friedman test was employed.

Pairwise Wilcoxon signed-rank tests were performed to determine which models were different when a statistically significant difference was found from the Friedman test. Three pairwise comparisons were made, and a Bonferroni correction was applied; therefore, the p-level was set at $p < .0167$. To illustrate and explain the main patterns from the quantitative results, qualitative examples were also chosen from the corpus.

RESULTS AND DISCUSSION

This section presents and analyzes the semantic accuracy and cultural equivalence of the translations of Jordanian proverbs produced by DeepSeek, Claude, and Grok. These results contain descriptive, inferential and qualitative analysis, and a comparison of the results across proverb categories.

Semantic Accuracy

In this section, the accuracy of the three AI models in capturing the intended meaning of the Jordanian proverbs is explored. Results are summarized by descriptive and inferential statistics, and then qualitative illustrations are included of the major patterns of semantic accuracy.

Descriptive Results of Semantic Accuracy Across the Three AI Models

The descriptive analysis examines the distribution of semantic accuracy scores across DeepSeek, Claude, and Grok. Table 3 presents the frequency and percentage of translations at each of the five semantic accuracy levels for the three models.

Table 3: Distribution of Semantic Accuracy Levels Across DeepSeek, Claude, and Grok

Score	Semantic Accuracy Level	AI Model	Frequency	Percentage (%)
		DeepSeek	126	66.32%
5	**Highly Accurate**	Claude	150	78.95%
		Grok	122	64.21%
		DeepSeek	41	21.58%
4	**Accurate**	Claude	29	15.26%
		Grok	45	23.68%
		DeepSeek	14	7.37%
3	**Moderately Accurate**	Claude	7	3.68%
		Grok	13	6.84%
		DeepSeek	6	3.16%
2	**Slightly Accurate**	Claude	3	1.58%
		Grok	8	4.21%
		DeepSeek	3	1.58%
1	**Inaccurate**	Claude	1	0.53%
		Grok	2	1.05%

As shown in Table 3, all three models achieved generally high levels of semantic accuracy, with most translations classified as Highly Accurate or Accurate. Claude recorded the highest proportion of Highly Accurate translations (78.95%), followed by DeepSeek (66.32%) and Grok (64.21%). In contrast, inaccurate translations were uncommon across all models, accounting for only 0.53% for Claude, 1.58% for DeepSeek, and 1.05% for Grok. Overall, the descriptive results indicate that Claude demonstrated the strongest semantic accuracy, while DeepSeek and Grok showed relatively similar patterns.

Inferential Comparison of Semantic Accuracy Across the Three AI Models

Inferential analysis was conducted to determine whether the differences in semantic accuracy among DeepSeek, Claude, and Grok were statistically significant. The Friedman test was first applied to compare the three models, followed by pairwise comparisons where appropriate.

Table 4: Friedman Test Results for Semantic Accuracy Across DeepSeek, Claude, and Grok

AI Model	N	Mean	SD	Mean Rank
DeepSeek	190	4.48	0.883	1.95
Claude	190	4.71	0.665	2.16
Grok	190	4.46	0.876	1.89
Friedman test	**190**			**$\chi^2(2) = 23.729, p < .001$**

Note. Scores ranged from 1 to 5 for all three models. Higher scores indicate greater semantic accuracy.

As shown in Table 4, Claude achieved the highest semantic accuracy (M = 4.71, SD = 0.665; Mean Rank = 2.16), followed by DeepSeek (M = 4.48; Mean Rank = 1.95) and Grok (M = 4.46; Mean Rank = 1.89). The Friedman test revealed a statistically significant difference among the three models, $\chi^2(2) = 23.729$, $p < .001$. Therefore, H01 was rejected, indicating that semantic accuracy differed significantly across the three AI models.

Since the Friedman test identified a significant overall difference, pairwise Wilcoxon signed-rank tests were conducted to determine which specific models differed. A Bonferroni-adjusted significance level of $\alpha = .0167$ was applied to the three comparisons.

Table 5: Pairwise Wilcoxon Signed-Rank Tests for Semantic Accuracy with Bonferroni Correction

Pairwise Comparison	First Model Higher	Second Model Higher	Ties	Z	p-value	Adjusted α	Result
DeepSeek vs Claude	18	45	127	−3.494	< .001	.0167	Significant; Claude higher
Claude vs Grok	47	14	129	−3.946	< .001	.0167	
DeepSeek vs Grok	22	15	153	−0.575	.565	.0167	Not significant

Note. Bonferroni-adjusted significance level: α = .05/3 = .0167. "Higher" represents the number of proverb translations for which one model received a higher semantic-accuracy score than the comparison model.

As presented in Table 5, Claude performed significantly better than DeepSeek (Z = −3.494, p < .001) and Grok (Z = −3.946, p < .001). In contrast, the difference between DeepSeek and Grok was not statistically significant (Z = −0.575, p = .565). Thus, the pairwise analysis indicates that the significant overall difference was mainly attributable to Claude's stronger semantic accuracy, while DeepSeek and Grok demonstrated statistically comparable performance.

The findings are consistent with previous research showing that semantic performance can vary considerably across AI translation models. Al-Jarf (2025a), for example, reported differences among AI and machine translation systems in rendering Arabic metaphorical expressions, with literal translation and difficulty interpreting figurative meaning contributing to semantic errors. Similarly, Paphonphat (2026) found substantial differences among large language models when translating idiomatic expressions, confirming that model choice can influence the accuracy of meaning transfer. The present study extends these findings to Jordanian proverbs, a linguistically and culturally specific form of figurative language. More importantly, it provides direct statistical evidence that Claude achieved significantly higher semantic accuracy than both DeepSeek and Grok, whereas no significant difference was found between DeepSeek and Grok. This demonstrates that even when the same proverbs and identical prompting conditions are used, AI models may differ significantly in their ability to preserve the intended meaning of Jordanian proverbial expressions.

Qualitative Analysis of Semantic Accuracy

To complement the quantitative findings, 15 representative translations were selected for qualitative analysis, with three examples drawn from each semantic accuracy level. The examples illustrate how the AI-generated translations ranged from full preservation of intended meaning to substantial semantic distortion or loss.

Table 6: Selected Examples of Semantic Accuracy Across the Five Evaluation Levels

Score	Semantic Accuracy Level	Jordanian Proverb	AI Model	English Translation
5	Highly Accurate	أرسلت حسن يجيب حسين، راحوا ولا جوني الاثنين.	Claude	I sent Hassan to fetch Hussein, and neither of them ever came back.
		قلب المؤمن دليله.	Grok	The believer's heart is his guide.
		اللي ما يعرف الصقر يشويه.	DeepSeek	He who doesn't know the falcon roasts it.
4	Accurate	العيب من أهل العيب مش عيب.	DeepSeek	The flaw from the flawed is no flaw.
		فرعون فرعن من قلة ناس تضبه.	Grok	Pharaoh pharaoh-ed from the scarcity of those who would put him down.
		الطول طول نخلة، والعقل عقل صخلة.	Claude	Tall as a palm tree, but with the sense of a lamb.
3	Moderately Accurate	عمر القصير ما بوكل تين.	Grok	The short one's life will not eat figs.
		من سلمك مذبحه لا تذبحه.	Claude	He who hands you his slaughtering knife, don't slaughter him.
		صامت صامت، وإفطرت على بصلة.	DeepSeek	Silent, silent, and I broke my fast with an onion.
2	Slightly Accurate	إذا شفت الحافي، قول يا كافي.	DeepSeek	If you see the barefoot, say 'O Sufficient.'
		اللي تعرف ديته اقتله.	Grok	The one whose nature you know, kill him.
		ما بيضحك للرغيف السخن.	Claude	No one laughs in the face of a warm loaf of bread.
1	Inaccurate	مشاطرك على مرة أبوك.	Grok	Sharing poverty with your father's bitterness.
		طفر وقلة زفر.	Claude	A leap with little effort behind it.
		اللي بيفشخ أكبر من فشخته، بيوقع وبتنفك رقبته.	DeepSeek	He who farts bigger than his own fart will fall and break his neck.

The qualitative analysis, as seen in Table 6, provides a closer examination of how the three AI models preserved, modified, or lost the intended meanings of the Jordanian proverbs

across the five semantic accuracy levels. The selected examples show that semantic performance depended not simply on whether a translation was literal, but on whether the intended proverbial meaning remained accessible in English.

At the Highly Accurate level (Score 5), all three models successfully preserved the central meaning. Claude's "I sent Hassan to fetch Hussein, and neither of them ever came back." accurately communicates both the literal situation and the broader idea of an attempt to solve a problem that produces an undesirable outcome. Grok's "The believer's heart is his guide." retains the intended association between inner feeling and guidance, while DeepSeek's "He who doesn't know the falcon roasts it." preserves the message that failure to recognise value can lead to improper treatment. These examples demonstrate that close reproduction of source imagery can still achieve high semantic accuracy when the figurative relationship remains understandable. This qualifies previous findings that frequently associated literal translation with meaning loss in Arabic proverbs (Al Kayed et al., 2023; Al Rousan & Shatnawi, 2023).

At Score 4 (Accurate), the central meanings remained identifiable despite minor lexical or figurative limitations. DeepSeek's "The flaw from the flawed is no flaw." conveys the general evaluative message but weakens the meaning of "العيب". Grok's "Pharaoh pharaoh-ed from the scarcity of those who would put him down." preserves the idea that tyranny develops when restraint is absent, despite its awkward formulation. Similarly, Claude's "Tall as a palm tree, but with the sense of a lamb." maintains the contrast between physical stature and limited intelligence, although the original animal image is modified. These cases support Benyahia's (2025) observation that semantic meaning may remain recoverable even when AI output contains lexical or idiomatic limitations.

The Moderately Accurate examples (Score 3) reveal more noticeable semantic modification. Grok's "The short one's life will not eat figs." retains key lexical elements but fails to establish their intended relationship. Claude's "He who hands you his slaughtering knife, don't slaughter him." preserves the broader message of restraint but misinterprets an important source element. DeepSeek's "Silent, silent, and I broke my fast with an onion." similarly misreads "صامت" as "silent" rather than "fasting". Such cases illustrate how dialectal ambiguity can weaken otherwise recognisable meanings, consistent with Kadaoui et al. (2023) and Abu Guba and Abu Quba (2025), who identified dialectal Arabic as a continuing challenge for AI translation.

At the Slightly Accurate level (Score 2), only limited semantic correspondence remained. DeepSeek's "If you see the barefoot, say 'O Sufficient.'" retains source elements but fails to communicate the message of recognising one's blessings. Grok's "The one whose nature you know, kill him." misinterprets "دية" as "nature", removing a central semantic component. Claude's "No one laughs in the face of a warm loaf of bread." preserves individual lexical elements while substantially changing the intended characterisation. These examples reflect the lexical and semantic distortions reported by Al-Jarf (2025a, 2025b) in AI translations of metaphorical Arabic expressions.

Finally, the Inaccurate translations (Score 1) demonstrate almost complete failure to recover intended meaning. Grok produced "Sharing poverty with your father's bitterness.", introducing concepts unrelated to the source meaning. Claude rendered another dialectal proverb as "A leap with little effort behind it.", replacing its intended reference to poverty and deprivation with physical movement. DeepSeek's "He who farts bigger than his own fart will fall and break his neck." represents an especially clear lexical misinterpretation that eliminates the intended warning against exceeding one's abilities. These failures strongly

support Al-Azzam's (2018) and Alarsan's (2025) argument that Jordanian proverbial meaning cannot always be recovered from surface wording because it depends on dialectal, social, and contextual knowledge. Overall, the progression from Scores 5 to 1 shows that semantic accuracy declines primarily when AI models fail to move beyond lexical meaning and correctly interpret dialectal or figurative relationships, rather than because literal translation is inherently inaccurate.

Cultural Equivalence

In this section, the extent of the preservation and/or representation of the cultural meanings of the Jordanian proverbs by the three AI models is examined. Descriptive and inferential results are reported, and qualitative examples of the key patterns of cultural equivalence are offered.

Descriptive Results of Cultural Equivalence Across the Three AI Models

The descriptive analysis examines how cultural equivalence scores were distributed across DeepSeek, Claude, and Grok. Table 7 presents the frequency and percentage of translations at each of the five cultural equivalence levels for the three models.

Table 7: Distribution of Cultural Equivalence Levels Across DeepSeek, Claude, and Grok

Score	Cultural Equivalence Level	AI Model	Frequency	Percentage
5	**Highly Culturally Equivalent**	DeepSeek	73	38.42%
		Claude	118	62.11%
		Grok	66	34.74%
4	**Culturally Equivalent**	DeepSeek	54	28.42%
		Claude	47	24.74%
		Grok	63	33.16%
3	**Moderately Culturally Equivalent**	DeepSeek	40	21.05%
		Claude	14	7.37%
		Grok	38	20.00%
2	**Slightly Culturally Equivalent**	DeepSeek	16	8.42%
		Claude	8	4.21%
		Grok	14	7.37%
1	**Culturally Non-Equivalent**	DeepSeek	7	3.68%
		Claude	3	1.58%
		Grok	9	4.74%
Total		**DeepSeek**	**190**	**100.00%**
		Claude	**190**	**100.00%**
		Grok	**190**	**100.00%**

As presented in Table 7, the three models showed noticeable variation in their ability to preserve cultural equivalence. Claude achieved the highest proportion of Highly Culturally Equivalent translations (62.11%), compared with DeepSeek (38.42%) and Grok (34.74%). Lower levels of cultural equivalence were more frequent for DeepSeek and Grok, while Culturally Non-Equivalent translations remained relatively limited, representing 1.58% for Claude, 3.68% for DeepSeek, and 4.74% for Grok. Taken together, these results show a clear descriptive advantage for Claude in preserving cultural meaning, whereas DeepSeek and Grok displayed broadly comparable distributions.

Inferential Comparison of Cultural Equivalence Across the Three AI Models

To examine whether the observed differences in cultural equivalence were statistically meaningful, the three AI models were compared using the Friedman test. Pairwise Wilcoxon signed-rank tests were subsequently used to identify the specific differences between models when the overall comparison was significant.

Table 8: Friedman Test Results for Cultural Equivalence Across DeepSeek, Claude, and Grok

AI Model	N	Mean	SD	Mean Rank
DeepSeek	190	3.89	1.122	1.86
Claude	190	4.42	0.915	2.35
Grok	190	3.86	1.120	1.79
Friedman test	190			$\chi^2(2) = 79.831$, $p < .001$

Note. Scores ranged from 1 to 5 for all three models. Higher scores indicate greater cultural equivalence.

As reported in Table 8, Claude obtained the highest cultural equivalence score (M = 4.42, SD = 0.915; Mean Rank = 2.35), whereas DeepSeek (M = 3.89, SD = 1.122; Mean Rank = 1.86) and Grok (M = 3.86, SD = 1.120; Mean Rank = 1.79) achieved lower and relatively similar scores. The Friedman test confirmed a statistically significant difference in cultural equivalence across the three models, $\chi^2(2) = 79.831$, $p < .001$. Accordingly, H02 was rejected, demonstrating that the models differed significantly in their ability to preserve cultural equivalence.

Following this significant overall result, pairwise Wilcoxon signed-rank tests were performed using the Bonferroni-adjusted significance level of $\alpha = .0167$. These comparisons were used to determine whether the difference involved all three models or was primarily associated with the performance of a particular model.

Table 9: Pairwise Wilcoxon Signed-Rank Tests for Cultural Equivalence with Bonferroni Correction

Pairwise Comparison	First Model Higher	Second Model Higher	Ties	Z	p-value	Adjusted α	Result
DeepSeek vs Claude	13	75	102	−6.101	< .001	.0167	Significant; Claude higher
Claude vs Grok	79	9	102	−6.462	< .001	.0167	Significant; Claude higher
DeepSeek vs Grok	29	21	140	−0.654	.513	.0167	Not significant

Note. Bonferroni-adjusted significance level: a = .05/3 = .0167. "Higher" represents the number of proverb translations for which one model received a higher cultural-equivalence score than the comparison model.

As indicated in Table 9, Claude achieved significantly higher cultural equivalence than DeepSeek (Z = −6.101, p < .001) and Grok (Z = −6.462, p < .001). However, no statistically significant difference was observed between DeepSeek and Grok (Z = −0.654, p = .513). The pairwise findings therefore show that the overall difference was largely driven by Claude's stronger performance, while DeepSeek and Grok were statistically comparable in their preservation of cultural equivalence.

These findings reinforce previous evidence that AI systems differ in their capacity to handle culturally embedded and figurative language. Al-Kaabi et al. (2024) found that AI-generated translations could maintain linguistic fluency while losing important cultural imagery and connotations, while Khoshafah (2024) similarly reported difficulties in preserving culturally specific imagery and metaphorical meaning. Model-related variation has also been identified by Obeidat et al. (2024), whose findings showed substantial differences among translation systems in their ability to recognise and convey idiomatic meaning. The present results extend this evidence to Jordanian proverbs by demonstrating that cultural equivalence varies significantly even when the models translate the same source expressions under identical prompting conditions. In this corpus, Claude showed a clear advantage over DeepSeek and Grok in conveying the culturally embedded meanings of Jordanian proverbial expressions.

Qualitative Analysis of Cultural Equivalence

To complement the quantitative findings, 15 representative translations were selected for qualitative analysis, with three examples drawn from each cultural equivalence level. The examples illustrate different degrees of cultural preservation, ranging from successful transfer of culturally embedded meanings and associations to substantial cultural loss or misrepresentation.

Table 10: Selected Examples of Cultural Equivalence Across the Five Evaluation Levels

Score	Cultural Equivalence Level	Jordanian Proverb	AI Model	English Translation
5	Highly Culturally Equivalent	قلب المؤمن دليله.	DeepSeek	The believer's heart is his guide.
		السبع سبع، لو حطيته بقفص.	Grok	A lion is a lion, even if you put him in a cage.
		كلب بنبح ما بعض.	Claude	A barking dog doesn't bite.
4	Culturally Equivalent	أرسلت حسن يجيب حسين، راحوا ولا جوني الاثنين.	DeepSeek	I sent Hassan to fetch Hussein; they both left and neither came back.
		ألف عين تبكي، ولا عين أمي تدمع.	Grok	A thousand eyes may cry, but not my mother's eye shall tear.
		يا داخل بين البصلة وقشرتها، ما ينوبك إلا ريحتها.	Claude	He who comes between the onion and its skin gets nothing but its smell.
3	Moderately Culturally Equivalent	العيب من أهل العيب مش عيب.	DeepSeek	The flaw from the flawed is no flaw.
		صاحب الحق عينه قوية.	Grok	The one who is in the right, his eye is strong.

		مثل صابونة الميت، منين ما قبضتها بتمزط.	Claude	Like a dead man's soap, wherever you grasp it, it slips away.
		بقاتل ذبان وجهه.	DeepSeek	His face kills flies.
2	**Slightly Culturally Equivalent**	بورجيك نجوم الظهر.	Grok	I'll show you the stars at noon.
		اذبح بسّك ليلة عرسك.	Claude	Slaughter your cat on your wedding night.
		مثل خبز الصاج، انقلب على الوجهين.	DeepSeek	Two strikes to the head cause pain.
1	**Culturally Non-Equivalent**	طفر وقلة زفر.	Claude	A leap with little effort behind it.
		بخت الرمايم قايم، وبخت العدايل مايل.	Grok	The fortune of the leftovers is standing, and the fortune of the fresh is tilted.

The qualitative analysis further demonstrates how the three AI models differed in preserving the cultural meanings, associations, and imagery embedded in Jordanian proverbs, as presented in Table 10. The selected examples show a gradual movement from successful cultural transfer to translations in which culturally significant meanings were substantially weakened or completely lost.

At the Highly Culturally Equivalent level (Score 5), the models successfully retained culturally meaningful imagery while making the intended message accessible in English. DeepSeek's "The believer's heart is his guide." preserves the culturally familiar association between the heart, belief, and inner guidance. Grok's "A lion is a lion, even if you put him in a cage." retains the lion as a symbol of strength and inherent character despite external restriction. Claude's "A barking dog doesn't bite." is particularly successful because the expression is also familiar and readily interpretable in English, allowing both meaning and proverbial function to be maintained. These examples support Hamdi et al. (2023), who emphasised that cultural equivalence is possible when comparable proverbial meanings and associations can be established across languages.

At Score 4 (Culturally Equivalent), the main cultural messages were preserved despite minor differences in resonance or naturalness. DeepSeek's "I sent Hassan to fetch Hussein; they both left and neither came back." retains the culturally recognisable names and narrative structure through which the proverb communicates an unsuccessful attempt to obtain help. Grok's "A thousand eyes may cry, but not my mother's eye shall tear." preserves the strong cultural valuation of the mother and the emotional preference for protecting her from suffering. Claude's "He who comes between the onion and its skin gets nothing but its smell." similarly maintains the culturally embedded imagery warning against interfering in intimate relationships. Such examples illustrate Alarsan's (2025) argument that Jordanian proverbs carry social values and sociopragmatic meanings beyond their individual words.

At the Moderately Culturally Equivalent level (Score 3), the original cultural imagery remained visible, but its significance became less accessible to an English reader. DeepSeek's "The flaw from the flawed is no flaw." communicates a general evaluative message but weakens the socially embedded meaning of "العيب". Grok's "The one who is in the right, his eye is strong." reproduces the culturally meaningful image of a "strong eye", yet this metaphor is unlikely to communicate its

association with confidence derived from being right. Likewise, Claude's "Like a dead man's soap, wherever you grasp it, it slips away." preserves unusual source imagery without adequately transferring its cultural associations. This pattern reflects Maalej's (2009) finding that retaining surface imagery alone may not preserve the cultural schemas underlying Arabic proverbs.

At Score 2 (Slightly Culturally Equivalent), literal imagery was retained but much of its culturally intended function was inaccessible. DeepSeek's "His face kills flies.", Grok's "I'll show you the stars at noon.", and Claude's "Slaughter your cat on your wedding night." reproduce striking source images, but an English reader without Jordanian cultural knowledge would have difficulty recovering their intended social or pragmatic meanings. This illustrates the distinction between preserving cultural imagery and achieving cultural equivalence. Similar limitations were reported by Al-Kaabi et al. (2024) and Mohsen (2024), who found that AI systems could preserve surface elements while losing cultural connotations and symbolic depth.

Finally, the Culturally Non-Equivalent examples (Score 1) demonstrate substantial cultural loss or misrepresentation. DeepSeek's "Two strikes to the head cause pain." no longer preserves the original cultural image or message. Claude's "A leap with little effort behind it." results from misinterpreting dialectal wording and consequently removes its culturally embedded meaning. Grok's "The fortune of the leftovers is standing, and the fortune of the fresh is tilted." retains lexical traces of the source but does not communicate its culturally understood social meaning. These cases strongly reflect Al-Azzam's (2018) finding that Jordanian proverbs are rooted in social, historical, and cultural contexts that literal or decontextualised translation may fail to recover. Overall, the examples demonstrate that cultural equivalence depends not merely on retaining source imagery but on successfully communicating the cultural associations and pragmatic meanings carried by that imagery.

CONCLUSIONS

This study demonstrates that the ability of AI models to translate Jordanian proverbs cannot be judged solely by whether they reproduce the general meaning of the source expression. Although DeepSeek, Claude, and Grok were generally successful in conveying semantic content, preserving the cultural meaning of the same proverbs was more challenging. The central conclusion is therefore that semantic accuracy and cultural equivalence represent related but distinct dimensions of AI translation quality. A translation can be semantically understandable and largely accurate while still losing part of the cultural knowledge, social associations, or pragmatic meaning carried by the original proverb.

The comparison among the three models further shows that AI translation quality is model-dependent rather than uniform. Claude demonstrated the strongest overall performance in both semantic accuracy and cultural equivalence, whereas DeepSeek and Grok showed statistically comparable performance. More importantly, the difference between the models became more visible when cultural equivalence was considered. This suggests that the ability to convey basic or intended meaning does not necessarily reflect the same ability to interpret culturally embedded language. Evaluating AI translation only through linguistic or semantic accuracy may therefore provide an incomplete picture of model performance.

The qualitative findings help explain why this distinction matters. Literal translation itself was not necessarily problematic. Several close or literal translations successfully preserved the intended meaning because the original imagery remained understandable in English. The problem emerged when the models reproduced source words or images without recognising what they represented within Jordanian culture.

Expressions involving culturally meaningful images could remain linguistically visible in English while their underlying social or pragmatic meanings became inaccessible. Consequently, the effectiveness of literal translation depends on whether the relationship between the source image and its intended meaning can survive the movement between the two languages and cultures.

This finding is particularly important for Jordanian proverbs because their meanings are not contained entirely in their individual words. They draw on shared cultural knowledge, dialectal usage, social experience, figurative associations, and conventional understandings within the Jordanian community. The lower-performing examples showed that when a model misinterpreted a dialectal word or failed to recognise such an association, the resulting problem extended beyond lexical error. It could change the message of the proverb or preserve an image whose cultural function was no longer understandable.

The findings also provide support for applying Nida's theory of equivalence to contemporary AI translation. The results illustrate why correspondence between source and target wording is insufficient as the sole indication of successful translation. What ultimately matters is whether the target expression communicates the intended message and enables the target reader to access culturally significant meaning. In this respect, the study extends equivalence-based thinking from human translation to AI-generated translation and shows its continuing relevance for evaluating culturally embedded language.

The study contributes to AI translation research by demonstrating the value of evaluating semantic accuracy and cultural equivalence separately rather than treating translation quality as a single general construct. The distinction revealed information that would have been obscured by an overall accuracy score: the models were generally more successful at transferring meaning than at preserving cultural significance. The findings therefore suggest that future evaluation frameworks for AI translation should incorporate cultural and pragmatic dimensions alongside semantic accuracy, particularly when dealing with dialects, idioms, proverbs, and other culturally dependent expressions.

From a practical perspective, the results indicate that fluent or apparently accurate AI output should not automatically be accepted as culturally equivalent. Translators, researchers, and other users of AI translation need to evaluate whether culturally embedded meanings remain accessible to the intended audience. For AI developers, improving proverb translation requires greater exposure to dialectal Arabic and culturally contextualised data rather than merely increasing lexical correspondence between Arabic and English.

The study is limited to three AI models, one translation direction, and Jordanian proverbs generated under a specific prompting condition and period. Future research should therefore examine whether the same semantic–cultural gap appears across other Arabic dialects, language pairs, AI models, and culturally embedded forms of language. It would also be valuable to investigate whether providing cultural context or culturally enriched prompts can reduce this gap. Ultimately, progress in AI translation should not be measured only by whether a model can translate what a proverb says, but also by whether it can communicate what the proverb means within the culture that uses it.

References

Abdelhalim, S. M., Alsahil, A. A., & Alsuhaibani, Z. A. (2025). Artificial intelligence tools and literary translation: A comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives. *Cogent Arts & Humanities, 12*(1), 1–20.

https://doi.org/10.1080/23311983.2025.2508031

Abu Guba, M. N., & Abu Quba, A. (2025). AI translation: Evaluating ChatGPT's reliability in translating Arabic to English. *Journal of Artificial Intelligence and Technology, 5*, 551–556. https://doi.org/10.37965/jait.2025.0831

Ahmed, O. F., Dewi, I. K., & Anis, M. Y. (2025). Evaluation of machine translation systems: A literature review on ChatGPT and Google Translate. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, 13*(1), 106–125. https://doi.org/10.24256/ideas.v13i1.6236

Al-Azzam, B. H. S. (2018). Culture as a problem in the translation of Jordanian proverbs into English. *International Journal of Applied Linguistics & English Literature, 7*(1), 56–63.

Al-Jarf, R. (2025a). DeepSeek, Google Translate and Copilot's translation of Arabic grammatical terms used metaphorically. *Journal of Computer Science and Technology Studies, 7*(3), 46–57. https://doi.org/10.32996/jcsts

Al-Jarf, R. (2025b). Translation of Arabic expressions of impossibility by AI and student-translators: A comparative study. *Journal of Computer Science and Technology Studies, 7*(8), 288–299. https://doi.org/10.32996/jcsts

Al-Kaabi, M. H., AlQbailat, N. M., Badah, A., Ismail, I. A., & Hicham, K. B. (2024). Examining the cultural connotations in human and machine translations: A corpus study of Naguib Mahfouz's *Zuqāq al-Midaqq*. *Journal of Language Teaching and Research, 15*(3), 707–718. https://doi.org/10.17507/jltr.1503.03

Al Kayed, M., Alkayid, M., & Essa, L. B. (2023). A contrastive study of the connotative meanings of "dog-related" expressions in English and Jordanian proverbs: Implications for translators and language teachers. *Acta Linguistica Petropolitana, 19*(1), 66–101. https://doi.org/10.30842/alp2306573719166101

Alarsan, F., & Khan, T. (2025). Multi-word expressions in Jordanian Arabic: A socio-pragmatic study of idioms and proverbs. *International Journal of Language and Literary Studies, 7*(3), 260–277. https://doi.org/10.36892/ijlls.v7i3.2151

Aldawsari, A. H. (2024). Evaluating translation tools: Google Translate, Bing Translator, and Bing AI on Arabic colloquialisms. *Arab World English Journal (AWEJ) Special Issue on ChatGPT*, 237–251. https://doi.org/10.24093/awej/ChatGPT.16

Alharbi, A. S. (2023). Translating Saudi Najdi dialect proverbs into English: Challenges and strategies for preserving cultural meaning. *International Journal of Linguistics, Literature and Translation, 6*(12), 213–227. https://doi.org/10.32996/ijllt

Alhumsi, H., Belhassen, S., Alzobidy, S., & Hamda, A. (2025). Challenges in translating English phrasal verbs: A gender case of EFL undergraduate students at Saudi Electronic University. *Forum for Linguistic Studies, 7*(7), 134–145.

Ali, M. A. (2016). Artificial intelligence and natural language processing: The Arabic corpora in online translation software. *International Journal of Advanced and Applied Sciences, 3*(9), 59–66. https://doi.org/10.21833/ijaas.2016.09.010

Benyahia, W. (2025). On the assessment of ChatGPT-4 translation accuracy between Arabic and English with special reference to polysemous and culture-bound items. *Almodawana, 12*(1), 517–540.

Cahyaningrum, I. O. (2024). ChatGPT vs. Google Translate for translation. *The 3rd International Symposium on the Practice of Coexistence in Islamic Culture*.

de Waard, J., & Nida, E. A. (1986). *From one language to another: Functional equivalence in Bible translating*. Nelson.

Hamdi, S., Hashem, R., Holbah, W., Azi, Y., & Mohammed, S. (2023). Proverbs translation for intercultural interaction: A comparative study between Arabic and English using artificial intelligence. *World Journal of English Language, 13*(7), 282–291. https://doi.org/10.5430/wjel.v13n7p282

Kadaoui, K., Magdy, S. M., Waheed, A., Khondaker, M. T. I., El-Shangiti, A. O., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Tarjamat: Evaluation of Bard and ChatGPT on machine translation of ten

Arabic varieties. *arXiv*. https://arxiv.org/abs/2308.03051

Kadhim, P. (2024). The effect of using AI and Google Translate on translating BBC media texts into Arabic. *Al-Noor Journal for Digital Media Studies, 1*(2), 32–57. https://doi.org/10.69513/jnfdms.v.1.i.0.en.2

Khoshafah, M. A. N. A. (2024). Problems of ChatGPT in translating Yemeni Ibbi dialect into English. *Al-Qalam Journal of Humanities and Applied Sciences, 11*(44), 535–560.

Maalej, Z. (2009). A cognitive-pragmatic perspective on proverbs and its implications for translation. *International Journal of Arabic-English Studies, 10*(1), 135–154.

Mieder, W. (2004). *Proverbs: A handbook*. Greenwood Press.

Mohsen, M. (2024). Artificial intelligence in academic translation: A comparative study of large language models and Google Translate. *Psycholinguistics, 35*(2), 134–156. https://doi.org/10.31470/2309-1797-2024-35-2-134-156

Naveen, P., & Trojovsky, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience, 27*(10), 1–25. https://doi.org/10.1016/j.isci.2024.110878

Nida, E. A. (1964). *Toward a science of translating: With special reference to principles and procedures involved in Bible translating*. Brill Archive.

Nida, E. A., & Taber, C. R. (2003). *The theory and practice of translation*. Brill.

Obeidat, M., Haider, A., Tair, S., & Sahari, Y. (2024). Analyzing the performance of Gemini, ChatGPT, and Google Translate in rendering English idioms into Arabic. *FWU Journal of Social Sciences, 18*(4), 1–18. https://doi.org/10.51709/19951272/Winter2024/1

Paphonphat, K. (2026). AI-based translation of Korean idioms into Thai: Accuracy and error patterns. *Scientific Culture, 12*(1, Part 1), 525. https://doi.org/10.5281/zenodo.11425146

Saeed, A. M. A. (2025). Machine translation evaluation between Arabic and English during 2020 to 2024: A review study. *Arts for Linguistic & Literary Studies, 7*(2), 665–678. https://doi.org/10.53286/arts.v7i2.2528