

Original Research

AI-powered translation of Arabic idioms into English

Mohd Nour Al Salem^{1*}, Nimer Abusalim¹, Sharif Alghazo^{2,1}

¹The University of Jordan, Amman, Jordan

²University of Sharjah, Sharjah, United Arab Emirates

The study investigates how ChatGPT translates Arabic idioms into English, focusing on translation strategies, error types, and degrees of figurative preservation. A purposive corpus of 26 Arabic idioms was compiled, evenly divided between 13 transparent and 13 opaque expressions following Moon's distinction, and verified against lexicographic sources to ensure their conventional meaning and cultural authenticity. Translations were produced under two prompt conditions: a baseline instruction ('Translate into English') and an idiom-aware instruction ('Translate the following idiom into English'). Outputs were analysed qualitatively using a strategy-error framework evaluated through a three-point accuracy scale distinguishing full, partial, and inaccurate translations. The results suggest that transparent idioms are generally rendered successfully through established English equivalents, including under baseline prompting. Opaque idioms display greater variability: baseline prompts produce literal or paraphrastic renderings, whereas idiom-aware prompting increases idiomatic substitution and reduces literal translation. In many cases, idiomatic substitutions replace culturally situated imagery with conventional target-language idioms. The analysis demonstrates that prompt sensitivity interacts with idiom transparency and that semantic adequacy does not necessarily entail figurative richness. The study contributes evidence of the interaction between prompt design and idiom transparency, analysis of the systematic replacement of culturally situated imagery in LLM translations, and a qualitative evaluation framework for idiom translation in AI systems. The findings clarify how LLMs handle non-compositional expressions and identify conditions under which human oversight remains necessary in idiom translation tasks.

Keywords: ChatGPT, Arabic-English translation, idiom translation, translation quality, AI translation, AI prompt



This is an open access article distributed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/) (CC BY-NC 4.0), which allows its unrestricted use for non-commercial purposes, subject to attribution. The material can be shared/adapted for non-commercial purposes if you give appropriate credit, provide a link to the license, and indicate if changes were made.

1. INTRODUCTION

Idiomatic expressions play a key role in natural languages, as they convey cultural values, figurative meanings, and socio-historical experiences. They are widely recognised as fixed multiword expressions whose meanings cannot be deduced from the literal interpretations of their constituent words (Sadigzade, 2025). As Alarsan and Khan (2025) maintain, idioms transcend literal definitions insofar as they convey cultural meanings, historical contexts,

and shared community experiences. They function as condensed cultural signifiers that reflect the values and lived experiences of a speech community. Similarly, Makhmudova and Khamitov (2025) stress that idioms embody the core values and collective wisdom of the societies that produce them. These observations point to the dual linguistic and cultural nature of idioms that function as lexical units while also serving as socio-cultural expressions. Scholars have long emphasised that idioms pose challenges

Article history:

Submitted October 12, 2025

Revised February 27, 2026

Accepted March 16, 2026

CRediT Author Statement:

Mohd Nour Al Salem: Conceptualisation, Methodology, Writing – Original Draft.

Nimer Abusalim: Formal analysis, Investigation.

Sharif Alghazo: Writing – Review & Editing.

Conflict of interest:

The authors declared no conflict of interest.

Data availability statement:

The data supporting this study's findings are included within the article and/or its supplementary materials.

Funding:

No funding was reported for this research.

Doi:

10.22363/2521-442X-2026-10-1-8-21

For citation:

Al Salem, M. N., Abusalim, N., & Alghazo, S. (2026). AI-powered translation of Arabic idioms into English. *Training, Language and Culture*, 10(1), 8–21.

for second language learners, translators, and computational systems because of their non-compositional meaning. Moon (1998) provided a foundational typology by distinguishing between *transparent* and *opaque* idioms. Transparent idioms are those in which the figurative sense can be at least partially inferred from the literal image (e.g., *spill the beans*), whereas opaque idioms resist such inference and require prior cultural knowledge (e.g., *kick the bucket*). This distinction is crucial not only in translation studies but also in computational linguistics, since the degree of transparency often predicts the level of translation difficulty (Fadaee et al., 2018).

Arabic presents a particularly rich and diverse idiomatic repertoire, with many idioms deriving from Qur'anic imagery, Hadith, classical poetry, and regional proverb traditions (Ghazala, 2008; Hinds & Badawi, 1986). They may be colloquial or formal, with figurative images reflecting local values, social roles, and religious symbolism. For instance, idioms such as *عاد بخفي حنين* (lit. *returned with Hunayn's sandals*) convey the meaning of returning empty-handed but are steeped in a cultural tale not accessible to non-Arabic speakers. Likewise, *ألقى الحبل على الغارب* (lit. *he threw the rope on the withers*) figuratively means *letting things go uncontrolled*, but its imagery originates in Bedouin practices of tying camels. Idioms are thus repositories of culture-bound references.

The difficulty of translating idioms between Arabic and English stems from stark differences in linguistic structure, cultural imagery, and metaphorical conceptualisation. Adelnia and Dastjerdi (2011) and Baker (2018) argue that one of the most complex tasks for translators is reproducing idiomatic meaning with comparable connotations. For Baker (2018), successful idiom translation requires sensitivity to both linguistic principles and cultural expectations across languages. In Arabic–English translation pedagogy, idioms have long been treated as culture-bound expressions that pose a specific challenge for translators (Ghazala, 2008), for idiomatic transfer as a discourse-level operation requires attention to cultural factors (Hatim & Mason, 1990) and socio-cultural meaning (Newmark, 1988).

The advent of machine translation (MT) raised hopes that idioms could be rendered automatically across languages. However, MT has consistently struggled with idiomatic expressions. Early systems, largely rule-based, rendered idioms literally and often nonsensically. Neural Machine Translation (NMT), which uses deep learning over large parallel corpora, markedly improved fluency but continued to mistranslate idioms because it treated them compositionally. For example, Baziotis et al. (2022) demonstrated that NMT tended to produce literal translations of idioms, missing their figurative meanings. Yet they also found that pretraining on large-scale monolingual corpora improved idiom handling, suggesting that

contextual exposure supports figurative recognition even without direct idiom-specific training. Complementary computational analyses have further shown why literalism persists. Dankers et al. (2022) report that Transformer-based models, the architecture underlying NMT, process idioms 'too compositionally', failing to recognise them as holistic units. This structural bias leads to calques and word-for-word renderings, which researchers have sought to mitigate through idiom-aware training, lexicon injection, and treating idioms as multiword expressions (MWEs) (Ramisch, 2015; Zaninello & Birch, 2020). One promising approach reframes idioms as MWEs that require specialised modelling. Zaninello and Birch (2020) argue that explicitly adapting MT architectures to handle MWEs significantly improves idiom translation performance, with Ramisch (2015) also stressing that idioms should be modelled as lexicalised units and not just compositional phrases. This treatment of idioms as multiword expressions has informed recent work on MT evaluation and model design.

While the translation of idioms has been examined in various languages, relatively few studies address Arabic–English idioms in large-scale LLM systems. Much of the existing Arabic-focused work evaluates classical MT engines and hybrid AI systems, with limited attention to contemporary LLMs such as ChatGPT (Almahasees, 2021; Almaaytah, 2022; Darwish et al., 2025; Liu et al., 2023). Because Arabic idioms carry distinctive cultural and historical references, systematic qualitative study of their translation into English in contemporary LLMs is methodologically demanding, and such work remains limited.

To address this gap, the present study offers an exploratory analysis of how ChatGPT translates Arabic idioms into English, examining translation strategies, error types, and the accuracy with which figurative meaning is rendered. The analysis adopts Moon's (1998) transparent and opaque typology, allowing a systematic comparison of idioms with varying degrees of semantic transparency. A corpus of 26 idioms was compiled, evenly divided between transparent and opaque types, with meanings confirmed in sources such as Ibn Manẓūr (1956) and Al-Zabīdī (2011), and classical proverb collections such as Al-Maydānī (2005) and Al-'Askarī (1964). Translations were generated using two prompt conditions: a baseline prompt ('Translate into English') and an idiom-aware prompt ('Translate the following idiom into English').

This study is guided by three central questions:

1. What are the strategies ChatGPT uses to translate idioms from Arabic to English?
2. What errors occur in ChatGPT's translations of idioms (e.g., literalism, calque formation, non-idiomatic equivalents, register mismatch)?
3. How accurately does ChatGPT render the figurative meaning of Arabic idioms in English?

The study thus aims to offer several exploratory contributions. First, it expands idiom translation research to a language pair with strong cultural specificity but relatively limited computational attention. Second, it examines the role of prompt design in AI translation performance, assessing whether idiom-aware instructions improve output quality. Third, it proposes a qualitative framework for analysing AI translations of idioms, identifying strengths, weaknesses, and persistent challenges. Finally, the study assesses the need for human oversight in AI-assisted translation, particularly when dealing with culture-bound figurative language.

2. THEORETICAL BACKGROUND

Translation of idiomatic expressions has consistently posed a challenge for both human and machine translators. By their nature, idioms resist straightforward equivalence, since their figurative sense and culture-bound references cannot be reliably conveyed through literal translation. Unlike single lexical items, idioms function as fixed, non-compositional units whose meanings must be interpreted at the level of the expression as a whole. This makes them particularly problematic in cross-linguistic contexts, where cultural imagery, historical allusion, and pragmatic intent may differ between source and target languages. This difficulty is especially evident in Arabic idiomatic expressions, which are often rich in metaphor, intertextuality, and religious or socio-historical imagery, features that rarely have immediate or natural counterparts in English (Baker, 2018; Ghazala, 2008). As Baker (2018) observes, the challenge is finding target-language equivalents that preserve both semantic meaning and cultural connotations.

Idioms are widely discussed within the study of formulaic language. Wray (2002) contends that formulaic sequences are ubiquitous in processing and production and therefore should not be treated as marginal in translation. As Ramisch (2015) also emphasises, idioms belong to the class of multiword expressions, which are common, complex, and computationally demanding, and are therefore particularly problematic for translation technologies. Moon (1998) observes that idioms must be assessed in relation to the texts and contexts in which they occur, because their entailments and evaluative purposes are highly dependent on discourse.

Similarly, Ghazala (2008) treats idioms as culture-specific expressions that resist direct transfer and require adaptive translation strategies. Newmark (1988, pp. 95–96) likewise argues that translation involves the transfer of cultural meaning alongside linguistic form. He also distinguishes between universal language and culture-specific language, noting that culturally situated words often create translation difficulties when no shared reference exists, and maintains that idioms present a particular kind of

problem for translators since they require attention to referential and aesthetic aims (Newmark, 1988, pp. 53–58). Hatim and Mason (1990) similarly describe idioms as closely linked to socio-cultural frameworks and argue that their translation involves lexical substitution together with informed judgement about cultural meaning. Translation evaluation frameworks also address this issue, with House's (2015) revised *Translation Quality Assessment* explicitly foregrounding cultural and situational appropriateness ('cultural filtering') and offering a qualitative complement to string-based metrics when assessing idiom translations.

This dual linguistic and cultural complexity accounts for the longstanding difficulty of translating idioms accurately and appropriately.

Recent advances in AI-driven language technologies, particularly neural machine translation (NMT) models trained on massive multilingual datasets, have raised expectations for more accurate and fluent idiom translation. These models have demonstrated gains in surface fluency, producing target-language outputs that appear grammatically correct and stylistically natural. However, researchers caution that such apparent fluency may mask the loss of cultural and figurative meaning, particularly in the treatment of idioms which, as non-compositional and culturally bound expressions, continue to expose the limitations of data-driven systems. Baziotis et al. (2022) provide systematic evidence of this problem by introducing an automatic metric for measuring literal translation errors. Their findings show that NMT systems tend to default to literal renderings of idioms, thereby stripping them of their figurative meaning. They also report that monolingual pretraining improves models' ability to recognise idioms as fixed expressions rather than compositional phrases, which mitigates such errors.

Literal translation nevertheless remains common in practice. A widely cited example is the rendering of the Arabic idiom *بلغ السيل الزبى* as *the flood reached the threshold*. While linguistically accurate on a literal level, this translation fails to convey the idiomatic sense of reaching the limit of patience, illustrating how surface accuracy may obscure a deeper failure to capture figurative meaning. Cases such as this point to the gap between fluency and cultural adequacy and remind researchers and practitioners that idiom translation remains a challenge even for state-of-the-art neural systems.

Consistent with this observation, research on machine translation consistently identifies idioms as a persistent challenge due to their non-compositional and culturally situated nature. Neural systems tend to process language compositionally, which leads to literal renderings of figurative expressions (Dankers et al., 2022; Fadaee et al., 2018). This tendency is particularly problematic for idioms, whose meanings cannot be inferred from their

constituent parts. Studies show that even advanced neural models tend to default to word-for-word translations unless idioms are explicitly recognised as holistic units or treated as multiword expressions (MWEs) (Ramisch, 2015; Zaninello & Birch, 2020)

In the case of Arabic–English translation, this difficulty is amplified by the cultural and historical references carried by idioms. Evaluations of major MT systems report consistent literalism and loss of figurative meaning in Arabic idiom translation (Almaaytah, 2022; Almahasees, 2021). For example, culturally situated expressions may be rendered through paraphrase, which strips away metaphorical imagery. Yousef (2024), analysing 55 Arabic idioms in literary translation, found that paraphrase is the dominant strategy when direct equivalence is unavailable, confirming earlier claims that Arabic idioms require cultural adaptation (Baker, 2018; Ghazala, 2008).

This suggests that Arabic idioms may present a particularly useful test case for evaluating AI translation systems.

Recent work on large language models (LLMs) introduces a further factor of prompt sensitivity. Studies report that translation quality in LLMs is strongly influenced by task framing and instruction design (Yamada, 2023; Zhang et al., 2023; Zhu et al., 2024). In idiom-specific experiments, prompt modifications have been shown to increase the likelihood of idiomatic renderings instead of literal translations (Castaldo & Monti, 2024).

However, most existing studies focus on European language pairs or large-scale benchmarks. Systematic study of Arabic idioms in LLM translation remains limited, especially in controlled qualitative settings.

Evaluation of translation quality accordingly presents methodological challenges arising from the formal and functional complexity of idioms as culturally situated expressions. Automatic metrics such as BLEU cannot capture figurative adequacy or cultural connotations (Baziotis et al., 2022). Translation quality assessment frameworks instead emphasise semantic equivalence, register appropriateness, and cultural filtering (House, 2015). This supports qualitative, strategy-based analysis when figurative language is examined, particularly in small, controlled datasets.

Prior research thus establishes three key points relevant to the present study: (i) idioms are prone to literalism in neural translation; (ii) Arabic–English idioms pose additional cultural challenges; and (iii) LLM outputs are sensitive to prompt design.

What remains insufficiently studied is how these factors combine in a controlled Arabic idiom corpus analysed through the transparent vs. opaque typology. To address this gap, this study analyses 26 Arabic idioms under two prompt conditions, with attention to strategy use, error types, and degrees of figurative preservation.

3. METHODOLOGY

3.1. Design overview

This study evaluates ChatGPT’s translations of Arabic idioms using Moon’s (1998) transparent–opaque distinction as the primary typological lens. This typology provides a theoretically motivated basis for predicting translation difficulty: transparent idioms, whose figurative sense can be partly inferred from the literal image, are generally easier to translate than opaque idioms, whose meanings are not compositionally recoverable (cf. Baker, 2018; Fadaee et al., 2018).

The analysis focuses on three interrelated aspects: (i) the strategies ChatGPT uses, (ii) the errors it produces, and (iii) a qualitative judgment of accuracy for each idiom – i.e., whether the output communicates the conventional figurative meaning in pragmatically acceptable English. The study follows a qualitative descriptive design appropriate for close, text-sensitive evaluation of figurative language (Miles et al., 2014; Saldanha & O’Brien, 2014; Sandelowski, 2000). Instead of relying on heavy theoretical abstraction, qualitative description allows the analysis to remain close to the data, documenting what the system produces and how those outputs correspond to established translation-taxonomy labels (Baker, 1992, 2018; Newmark, 1988; Nida, 1964). Each idiom is therefore analysed in terms of strategy, error type, and overall accuracy, with emphasis on identifying recurring patterns and offering explanatory commentary grounded in translation theory and idiom research (House, 2015; Ramisch, 2015; Wray, 2002).

3.2. Data collection

A corpus of 26 idioms – balanced across 13 transparent and 13 opaque items – was compiled. The set includes idioms from Modern Standard Arabic (MSA) and widely used Jordanian Arabic expressions to cover both formal and colloquial usage. To ensure Arabic origin and conventional meanings, candidates were verified against classical lexica (Ibn Manẓūr, 1956); Al-Zabīdī, 2011), major proverb and idiom collections (Al-‘Askarī, 1964; Al-Maydānī, 2005), and dialect dictionaries for colloquial items (Hinds & Badawi, 1986). The triangulation minimises the risk of idiolectal or non-canonical meanings and follows established practices for constructing idiom test sets in translation evaluation (Fadaee et al., 2018; Zaninello & Birch, 2020).

The 26 idioms were selected through purposive sampling. The resulting dataset constitutes a theoretically balanced test set based on Moon’s (1998) transparency continuum and is not intended to represent corpus frequency distributions. The sample was evenly divided between transparent (n=13) and opaque (n=13) idioms to enable controlled comparison of translation behaviour across semantic types. Within each category, items were

chosen to ensure variation in (i) semantic fields (e.g., emotion, exaggeration, conflict, stubbornness, indifference), (ii) metaphorical source imagery (e.g., animal imagery, material objects, religious or historical references), and (iii) register (Modern Standard Arabic and widely used Jordanian Arabic expressions). Many of the selected idioms are commonly used in Arabic language teaching and widely cited in classical proverb collections and dictionaries. The primary selection criterion, however, was typological balance, not corpus frequency. The controlled diversity allows the dataset to serve as an analytical probe of LLM behaviour across culturally and semantically distinct idiomatic constructions.

It should be noted that the corpus size ($n=26$) is intentionally limited and does not aim at statistical generalisation. The dataset functions as a balanced test set for examining idiom transparency effects and is not intended as a frequency-based or large-scale benchmark.

3.3. Inclusion criteria

An expression was classified as an idiom if it met three criteria:

1. Form: a fixed multi-word expression (MWE) rather than a free combination.
2. Function: occurs within a sentence (not a free-standing proverb).
3. Meaning: exhibits figurative, non-compositional meaning that cannot be recovered through combination of its parts, to varying degrees of transparency (Moon, 1998; Ramisch, 2015).

These criteria follow the MWE perspective in linguistics and machine translation, which treats idioms as lexicalised units requiring special handling (Ramisch, 2015; Wray, 2002; Zaninello & Birch, 2020).

3.4. Transparency labels

Idioms were labelled along Moon's (1998) transparency continuum.

1. Transparent: the literal image offers a reasonable clue to the figurative meaning for a competent L1 reader (e.g., سلاح ذو حدين *a weapon with two edges* → *a double-edged sword*).
2. Opaque: the figurative meaning is not inferable from the literal image (e.g., عاد بـخُفَى حُنين *returned with Hunayn's sandals* → *returned empty-handed*).

This categorisation is widely used in idiom research and MT evaluation because it correlates with known processing and translation difficulty (Dankers et al., 2022; Fadaee et al., 2018).

3.5. Translation generation

Translations were generated using GPT-5 (model identifier: gpt-5) accessed via ChatGPT Plus (OpenAI web interface). Outputs were generated on 12–14 October

2025. The default ChatGPT interface configuration was used. No custom system prompt was added by the researchers. The standard platform-level system instructions available within the ChatGPT interface remained active but were not modified or supplemented. No additional contextual priming, few-shot examples, temperature adjustments, or role instructions were introduced.

Two prompts were used: (i) baseline — 'Translate into English'; (ii) idiom-aware — 'Translate the following idiom into English'. Each prompt was executed in a separate, newly opened chat session to minimise contextual carry-over.

Each idiom was translated once per prompt condition (baseline and idiom-aware), resulting in a total of 52 outputs (26 idioms × 2 conditions). No repeated sampling or regeneration was performed. Accordingly, the study records single-instance outputs and does not account for performance distributions.

Because large language models exhibit stochastic variability across runs, results should be interpreted as observations from a single controlled generation per condition. Future work may incorporate repeated-run sampling to assess output variability.

3.6. Translation strategies

Each output was coded for translation strategy using established taxonomies:

- (i) literal (word-for-word): direct, form-based rendering without figurative sense (Baker, 1992, 2018);
- (ii) equivalent idiom: an established target-language idiom conveying the same meaning (the form may differ) (Baker, 1992, 2018);
- (iii) paraphrase: plain, non-idiomatic wording conveying the figurative sense (Baker, 1992, 2018);
- (iv) explication: brief added information that makes cultural or conceptual content explicit (Nida, 1964);
- (v) loan/calque: transfer of source-language imagery into the target language, often producing unnatural phrasing (Baker, 1992, 2018);
- (vi) omission: idiomatic meaning partially or fully not conveyed (Baker, 1992, 2018);
- (vii) creative/cultural substitution: a different target-language set phrase or idiom that achieves a comparable pragmatic effect (Baker, 1992, 2018; Newmark, 1988).

The use of multiple widely cited frameworks situates the analysis within established translation-studies practice and allows comparison with previous research on idiom translation (House, 2015; Saldanha & O'Brien, 2014).

3.7. Errors

Outputs were additionally coded for error types adapted from translation studies and recent MT/LLM idiom research (Baziotis et al., 2022; Dankers et al., 2022; House, 2015; Newmark, 1988):

- (i) literalism/word-for-word: retention of surface imagery where a figurative rendering is required;
- (ii) calque formation: unnatural target-language structure mirroring source-language imagery;
- (iii) wrong or non-idiomatic equivalent: an English idiom (or set phrase) that fails to convey the source meaning or sounds forced in context;
- (iv) over-translation or under-translation: addition or omission of key semantic components of the idiom's meaning (scope, polarity, intensity, register).

The error scheme covers both form-faithful failures (literalism, calque) and sense-faithful failures (mismatch, scope errors) and corresponds to evaluation concerns in idiomatic MT (Baziotis et al., 2022; House, 2015). Where useful, judgments were considered in relation to cultural appropriateness ('cultural filtering') as recommended by House (2015), acknowledging that idiom quality involves pragmatic and socio-cultural adequacy, not only lexical choice.

3.8. Coding procedure, evaluation, and reliability

The coding process involved three analytical layers: (i) translation strategy classification, (ii) error-type identification, and (iii) overall accuracy rating.

Two researchers (the first and third authors) participated in the coding process. Both have formal training in translation studies and applied linguistics. The outputs were independently coded by both researchers for translation strategy (literal, equivalent idiom, paraphrase, calque, explication, omission, creative substitution) and (error type (literalism, calque formation, wrong or non-idiomatic equivalent, over- or under-translation)).

Coding was performed independently at the initial stage. Coders worked from a shared coding manual derived from Baker (1992, 2018), Newmark (1988), Nida (1964), and House (2015), with operational definitions and examples agreed upon prior to the start of coding. Coders were not blind to prompt condition, as the two prompt types were evident in the output logs; however, coding decisions were based solely on the linguistic characteristics of the outputs.

Initial independent coding produced agreement rates of 88% for strategy classification and 85% for error-type classification. Disagreements primarily involved borderline cases between paraphrase and equivalent idiom, and between semantic narrowing and partial accuracy.

All disagreements were resolved through joint discussion until consensus was reached. No third adjudicator was required.

Given the qualitative orientation of the study and the limited dataset size, agreement was calculated as percentage agreement without resorting to statistical coefficients (e.g., Cohen's kappa), which are less stable with small categorical datasets.

To operationalise the concept of *accuracy*, evaluation was guided by a five-category qualitative grid distinguishing semantic equivalence, idiomaticity, register alignment, and cultural imagery retention.

Each output was assigned one of the following primary categories:

A. Full idiomatic match: established target-language idiom; preserves core semantic meaning; maintains comparable figurative force; register appropriate; no significant semantic shift.

B. Accurate paraphrase: core figurative meaning conveyed correctly; no established target-language idiom used; figurative imagery neutralised; register acceptable.

C. Partial semantic shift: core meaning broadly preserved; noticeable narrowing, broadening, or causal shift; idiomatic substitution may alter conceptual focus.

D. Register distortion: meaning largely preserved; tone, intensity, or pragmatic force misaligned; over- or under-formality relative to the source idiom.

E. Cultural imagery loss: meaning accurately conveyed; source metaphor or culturally embedded imagery replaced; figurative richness reduced.

Categories (C)–(E) may overlap conceptually; however, coders assigned the dominant evaluative feature for analytical clarity. For reporting purposes, categories (A) and (B) were treated as semantically accurate, whereas (C)–(E) represent graded forms of partial deviation rather than outright mistranslation.

Agreement on accuracy ratings reached 90%. Disagreements were resolved through discussion and reference to authoritative lexical sources. Independent double-coding increases analytical reliability; however, the analysis remains qualitative and interpretive. The purpose is to identify systematic patterns and does not involve establishing statistically generalisable reliability coefficients.

The evaluation in this study consists of expert qualitative judgment conducted by researchers trained in translation studies and applied linguistics. The coding and accuracy assessments rely on disciplinary expertise in idiom semantics, translation strategy taxonomy, and cross-cultural equivalence. The objective is to provide theoretically grounded interpretive analysis of LLM outputs. Accordingly, all evaluative decisions are grounded in established translation frameworks and authoritative lexical sources rather than subjective preference.

3.9. Operational definition of figurative richness

Throughout the analysis, the term *figurative richness* refers to the degree to which a translation preserves the metaphorical structure, cultural imagery, and expressive force of the source idiom in addition to conveying its propositional meaning. For the purposes of this study, figurative richness is operationalised through three evaluative criteria:

1. Imagery retention: whether the translation maintains a metaphorical image (either the original image or an established target-language metaphor) or reduces the expression to descriptive paraphrase.

2. Cultural or symbolic resonance: whether the translation preserves culturally situated imagery or replaces it with an idiom carrying comparable cultural weight in the target language.

3. Expressive intensity and register conformity: whether the translation reproduces the pragmatic force (e.g., emotional strength, irony, exaggeration, or colloquial tone) of the original idiom.

A translation may therefore be semantically accurate while exhibiting reduced figurative richness when metaphorical or culturally loaded imagery is replaced with neutral paraphrase. This distinction follows House's (2015) emphasis on cultural filtering and Baker's (2018) differentiation between paraphrase and idiomatic substitution. Hence, richness is hereby treated as a component in the evaluation of figurative fidelity.

4. STUDY RESULTS

The dataset consists of 26 idioms, equally divided into transparent idioms (Idioms 1–13) and opaque idioms (Idioms 14–26), following Moon's (1998) distinction.

Each idiom was tested under two prompt conditions: a baseline prompt ('Translate into English') and an idiom-aware prompt ('Translate the following idiom into English').

For each item, the Arabic meaning was confirmed through classical and modern references, and the outputs were evaluated according to the three criteria of strategy (the method used by ChatGPT – literal, equivalent idiom, paraphrase, calque, explication, omission, or creative substitution), error type (whether the translation exhibited literalism, non-idiomatic equivalents, calques, or semantic loss), and accuracy judgment (whether the figurative sense and cultural adequacy were preserved).

For clarity of presentation, results are presented in two parts for transparent (Table 1) and opaque idioms (Table 2).

For transparent idioms, ChatGPT tends to perform well when the figurative meaning is easily inferable from the literal image and when well-established English equivalents exist. Idioms such as سلاح ذو حدين (*a weapon with two edges*), عمِلَ مِنَ الحَبَّةِ قُبَّةً (*made a dome from a grain*), and على قدم وساق (*on foot and leg*) are translated into conventional English idioms like *a double-edged sword*, *to make a mountain out of a molehill*, and *in full swing*. These renderings indicate reliance on established target-language idioms, supporting Baker's (2018) principle of selecting idioms with similar meaning but dissimilar form. Even when the Arabic imagery is not fully preserved – for example, شَدَّ الرَّحَالَ loses its original equestrian imagery in the English rendering –

ChatGPT still succeeds in conveying figurative meaning through fluent and idiomatic English. The results suggest that transparent idioms present fewer difficulties for AI systems, as their meanings match existing target-language idioms and therefore reduce the likelihood of literal translation.

At the same time, limitations are visible in specific cases. For انقلب السحر على الساحر (*the magic turned against the magician*), both baseline and idiom-aware outputs defaulted to the idiom *the tables turned*, an English expression that implies a reversal of advantage but not the sense of a back-fired scheme. This suggests that ChatGPT may select idioms based on surface similarity rather than semantic precision. Similarly, idioms such as أثلج صدرى reveal a divergence between paraphrase and idiomatic rendering, with baseline outputs producing literal or plain phrasing while idiom-aware prompts elicit more idiomatic equivalents. These cases suggest that, although transparent idioms are translated with relatively high accuracy, ChatGPT tends to favour conventional English expressions over culturally specific imagery, resulting in semantic adequacy but partial cultural loss. The pattern indicates that prompt design improves idiomaticity, while the selection of equivalents remains guided largely by distributional associations, not contextual interpretation.

The second set of results concerns opaque idioms (Idioms 14–26), where figurative meaning is not recoverable from the literal image (Table 2).

Performance on opaque idioms is more uneven, indicating the limits of AI systems when cultural knowledge or unfamiliar imagery is required. Idioms in this group are rooted in Arab history, folklore, or metaphorical conventions, such as عاد بخفي حنين (*returned with Hunayn's sandals*) or بين حانا ومانا (*between Hana and Mana*). In these cases, ChatGPT consistently omits the culturally situated imagery, opting instead for paraphrases or widely used English idioms. For instance, the rendering *he came back empty-handed* conveys the correct figurative sense of عاد بخفي حنين, but it removes the reference to Hunayn's sandals, which carries cultural significance in Arabic tradition. Similarly, بين حانا ومانا is rendered as *caught between a rock and a hard place*, a strong English idiom that replaces culturally specific names with a more general metaphor of constraint. These examples suggest that ChatGPT prioritises communicative adequacy in the target language, with culturally specific imagery frequently replaced by more familiar English expressions.

Nonetheless, idiom-aware prompting improves the quality of the outputs via favouring idiomatic equivalents over paraphrases. Instances like أكل بعقلها حلوة (*he ate sweets with her mind*) are a case in point: the baseline paraphrase *he easily deceived her* is semantically accurate, but the idiom-aware output *he wrapped her around his finger*

Table 1
 Transparent idioms

SOURCE IDIOM	LITERAL GLOSS	INTENDED MEANING	BASELINE OUTPUT	IDIOM-AWARE OUTPUT	B STRATEGY	IA STRATEGY	ACCURACY (B/IA)
سلاح ذو حدين	weapon with two edges	double consequence	double-edged sword	double-edged sword	equivalent idiom	equivalent idiom	3 / 3 conventional L2 parallel
فاض الكيل	measure overflowed	patience exhausted	enough is enough	enough is enough	equivalent idiom	equivalent idiom	3 / 3 strong alignment
أثلج صدري	cooled my chest	felt relief	I was comforted	warmed my heart	paraphrase	equivalent idiom	2 / 3 imagery shift
نجا بجلده	escaped with skin	escaped narrowly	by the skin of his teeth	same	equivalent idiom	equivalent idiom	3 / 3 biblical parallel
سال لعابه	saliva ran	strong desire	he drooled	his mouth watered	literal/paraphrase	equivalent idiom	2 / 3 idiomatic upgrade
اشتد ساعده	arm grew strong	became capable	he grew stronger	came into his own	literal	equivalent idiom	2 / 3 figurative enhancement
انقلب السحر على الساحر	magic turned	plan backfired	the tables turned	the tables turned	equivalent idiom	equivalent idiom	2 / 2 semantic narrowing (backfire → ..)
دس السم في العسل	poison in honey	concealed harm	sugarcoat poison	same	calque/substitution	same	3 / 3 imagery softened
شد الرحال	tightened saddles	set out	set off	hit the road	paraphrase	equivalent idiom	3 / 3 image loss
عمل من الحبة قبة	dome from grain	exaggeration	mountain out of molehill	same	equivalent idiom	equivalent idiom	3 / 3 strong alignment
ذهبت أذراج الرياح	went with winds	came to nothing	went with the wind	went up in smoke	literal echo	equivalent idiom	2 / 3 figurative strengthening
طاش صوابه	reason flew	lost control	lost his mind	same	equivalent idiom	equivalent idiom	3 / 3 accurate
على قدم وساق	foot and leg	in full activity	in full swing	same	equivalent idiom	equivalent idiom	3 / 3 accurate

provides a closer match in expressive tone and figurative depth. The same pattern appears in وقع في حيص بيص, where the idiom-aware output produced *he was in a pickle* instead of the plainer *he was in great confusion*. This suggests that opaque idioms benefit most from idiom-aware prompting. Yet even in these cases ChatGPT rarely preserves the original cultural imagery, preferring instead to express the figurative meaning through familiar English idioms.

In sum, opaque idioms expose the model's reliance on surface fluency and its weakness in cultural transfer: while the figurative sense is usually retained, the imagery, historical context, and cultural colour of the Arabic originals are systematically removed, which reinforces the need for human oversight when translating culturally bound expressions.

5. DISCUSSION

Arabic idioms were divided into transparent and opaque types, following Moon's (1998) classification. The analysis examined ChatGPT's translation strategies, identified errors, and evaluated the accuracy of the intended meaning. Based on the results above, three recurrent patterns can be identified. The first pattern occurs when ChatGPT renders the Arabic idiom as an idiom in the target language in both the baseline and idiom-aware prompts. For example, سلاح ذو حدين (*a double-edged weapon*) was rendered in both cases as a *double-edged sword*, which preserves the figurative meaning of something that has both benefits and risks. A similar result appears with فاض الكيل (*the measure overflowed*), rendered as *enough is enough*, conveying the intended meaning of the idiom in both out-

Table 2
Opaque idioms

IDIOM	LITERAL GLOSS	INTENDED MEANING	BASELINE OUTPUT	IDIOM-AWARE OUTPUT	B STRATEGY	IA STRATEGY	ACCURACY (B/IA)
عاد بخفي حنين	Hunayn's sandals	empty-handed	came back empty-handed	returned empty-handed	paraphrase	paraphrase	3 / 3 cultural imagery neutralised
بيت القصيد	line of poem	central point	heart of the matter	crux of the matter	equivalent idiom	equivalent idiom	3 / 3 poetic image loss
بين حانا ومانا	between Hana & Mana	caught between pressures	caught between two sides	rock and hard place	paraphrase	equivalent idiom	2 / 3 idiomatic upgrade
قلب له ظهر المجن	back of shield	turned hostile	turned against him	turned his back on him	paraphrase	equivalent idiom	3 / 3 partial image loss
وقع في حيص بيص	hays bays	confusion	great confusion	in a pickle	paraphrase	equivalent idiom	2 / 3 figurative
ألقى الحبل على الغارب	rope on withers	let go uncontrolled	let things go	let things slide	paraphrase	equivalent idiom	3 / 3 image loss
ألقى الكلام على عواهنه	words carelessly	spoke recklessly	spoke recklessly	shot his mouth off	paraphrase	equivalent idiom	3 / 3 register intensification
أذن من طين وعجين	clay/dough ear	ignored	turned a deaf ear	same	equivalent idiom	equivalent idiom	3 / 3 image substitution
ركب رأسه	mounted head	stubborn	was stubborn	wouldn't budge	paraphrase	equivalent idiom	3 / 3 figurative gain
نسيج وحده	weave of his own	unique	one of a kind	same	equivalent idiom	equivalent idiom	3 / 3 accurate
أكل بعقلها حلوة	ate sweets with mind	manipulated	easily deceived	wrapped around finger	paraphrase	equivalent idiom	2 / 3 figurative gain
لا ناقة لي ولا جمل	no camel	nothing to do with it	nothing to do with it	no stake in it	paraphrase	equivalent idiom	3 / 3 imagery replaced
رفع عقيرته	raised larynx	raised voice	raised his voice	same	paraphrase	paraphrase	3 / 3 figurative flattening

puts, as described in Arabic Language Academy (2004). The idiom انقلب السحر على الساحر (*the magic turned against the magician*) was also rendered in both cases as *the tables turned*, with only a minor grammatical variation that does not affect interpretation. However, the rendering of انقلب السحر على الساحر as *the tables turned* represents a more complex case. Although both expressions indicate a reversal of advantage, the Arabic idiom specifically encodes the idea that a scheme backfires on its originator. The English idiom, by contrast, refers to a reversal of power relations rather than the self-defeating outcome of a deliberate action. The translation therefore involves a semantic shift from 'back-fire' to 'reversal'. Accordingly, the translation should be classified not as fully equivalent but as partially accurate, since the causal element embedded in the Arabic expression is weakened. This example shows that idiomatic substitution may preserve surface plausibility while altering the semantic core of the source idiom. Other examples

include على قدم وساق (*on a foot and leg*) which was rendered as *in full swing*, accurately conveying a stage at which an activity is at its highest level (CD, 2026a). This English idiom corresponds to the meaning recorded in Ibn Manzūr (1956). Similarly, طاش صوابه (*his reason flew*), was rendered as the English idiom *he lost his mind*, which conveys the figurative meaning without literal translation; according to the Cambridge Dictionary, the expression refers to behaving in a strange or irrational way (CD, 2026c). A comparable case appears in عمل من الحبة قسبة (*made a dome from a grain*), translated as the English idiom *to make a mountain out of a molehill*, which corresponds to the meaning recorded in Al-Maydānī (2005). The idioms نجا بجلده (*he escaped with his skin*) and قُطع دابر القوم (*cut off the remnant*) were likewise rendered with idiomatic equivalents in the target language.

These examples indicate that ChatGPT is capable of producing natural English idioms that convey the meaning of the Arabic source expressions without resorting to

word-for-word translation. Verification against the Cambridge Dictionary confirms that the English renderings correspond to established idioms.

In the second pattern, ChatGPT rendered some idioms accurately in the idiom-aware condition, producing an equivalent idiom in the target language. However, in the baseline condition it sometimes produced a meaningful translation even though the result was not an established idiom. For example, the idiom *أشلت صدري* (*my chest was cooled*) was translated literally in the baseline prompt. Although the output did not function as a standard English idiom, the translation remained meaningful. In contrast, the idiom-aware prompt produced *it warmed my heart*, an idiomatic expression that conveys the figurative meaning.

A similar case appears with the idiom *أشئت ساعده* (*his arm grew strong*). With the baseline prompt, ChatGPT generated the natural translation *he grew stronger*, which conveys the intended idiomatic sense. However, the idiom-aware prompt rendered it as the idiom *he came into his own*. This example shows that idiom-aware prompting can yield a conventional English idiom, although the semantic focus may differ slightly from the original expression.

The third pattern occurs when ChatGPT produces a literal translation in the baseline condition but generates an idiomatic equivalent when the idiom-aware prompt is used. This appears with *سال لعابه* (*his saliva ran*). Using the idiom-aware prompt, ChatGPT rendered it as the standard English idiom *his mouth watered*, which conveys the same meaning as described in Arabic Language Academy (2004). Without the idiom-aware instruction, the baseline prompt produced a more literal rendering, accompanied by the explanation *he drooled with desire*. This result suggests that ChatGPT recognised the figurative meaning of the Arabic idiom and attempted to express the sense of desire, even though the output was not presented as a conventional English idiom.

A comparable example appears in *ذهبت أذراج الرياح* (*went with the winds*), which carries the meaning *came to nothing* in Al-Maydānī (2005). With the idiom-aware prompt, ChatGPT produced the idiom *went up in smoke*, conveying the intended figurative meaning. This expression is also recorded in the Cambridge Dictionary as an established English idiom meaning *to be destroyed* (CD, 2026b). In contrast, the baseline output relied on a more literal rendering while attempting to preserve the general meaning.

These examples indicate that explicit identification of an expression as an idiom increases the likelihood that ChatGPT will produce a conventional idiomatic equivalent in English. When the idiomatic nature of the expression is not indicated, the system tends to produce literal or explanatory translations rather than established idioms.

A clarification is necessary regarding the evaluative stance adopted in this study. The systematic replacement of culturally situated imagery (e.g., Hunayn's sandals, the

rope on the withers, the shield metaphor) with familiar English idioms is not treated here as a translation 'error' in the professional sense. In translation theory, communicative equivalence often takes precedence over image retention, particularly when the target readership does not share the source cultural background (Baker, 2018; Newmark, 1988). From this perspective, rendering *عاد بخفي حنين* as *returned empty-handed* constitutes an acceptable and professionally defensible translational decision. However, within the analytical framework of this study, such substitutions are marked as instances of cultural imagery loss, because the metaphorical and historical specificity of the source expression is neutralised. This designation is descriptive, not prescriptive. It does not imply translational failure; rather, it records a shift from source-culture metaphor retention to target-language communicative naturalness. Thus, imagery loss is interpreted as a trade-off rather than a deficiency: semantic adequacy is preserved, but the figurative and cultural texture of the original expression is reduced.

In the situations described above, no strategy errors or accuracy errors were identified, which indicates that ChatGPT is capable of translating Arabic idioms into common English idioms in many cases, particularly when dealing with transparent idioms. In some instances, minor inaccuracies were observed, but these did not substantially alter the intended meaning, even when an equally conventional English idiom was not produced.

In the case of opaque idioms, a greater degree of difficulty was initially expected, since such expressions cannot be translated literally without losing their figurative meaning. While ChatGPT preserved the core figurative sense in most opaque idioms, several cases involved semantic narrowing or conceptual change rather than full equivalence. These instances are therefore better classified as partially accurate renderings, because small but meaningful differences in meaning remain.

The data indicates a clear trend. In some cases, the baseline output rendered the idiom in a simple descriptive form, while the idiom-aware prompt produced a figurative equivalent in English. For example, *بين حان ومانا* (*between Hana and Mana*) was rendered with the idiom-aware prompt as *caught between a rock and a hard place*, preserving the figurative sense while conveying the intended meaning of being between two opposing pressures (Al-Maydānī, 2005). In contrast, the baseline output conveyed the general meaning in a simpler form, lacking the idiomatic richness of the original.

Similarly, *لا ناقة لي ولا جمل* was rendered in the baseline condition as *have nothing to do with it*, whereas the idiom-aware translation produced the idiomatic *I have no stake in it*. Other examples show the same pattern. The idiom *ركب رأسه* (*acted stubbornly*) was rendered in the idiom-aware condition as *he wouldn't budge*, a figurative expression stronger

'In the situations described above, no strategy errors or accuracy errors were identified, which indicates that ChatGPT is capable of translating Arabic idioms into common English idioms in many cases, particularly when dealing with transparent idioms. In some instances, minor inaccuracies were observed, but these did not substantially alter the intended meaning, even when an equally conventional English idiom was not produced. In the case of opaque idioms, a greater degree of difficulty was initially expected, since such expressions cannot be translated literally without losing their figurative meaning. While ChatGPT preserved the core figurative sense in most opaque idioms, several cases involved semantic narrowing or conceptual change rather than full equivalence.'

than the descriptive baseline *he was stubborn*. The idiom فَلَبَّ لَه ظَهْرَ الْمِجَنِّ (he turned to him the back of the shield) was rendered idiomatically as *turned hostile against someone*, while the baseline produced *he turned his back on him*, which conveys the meaning but reduces the figurative nuance. Likewise, أَلْقَى حَبْلَهُ عَلَى الْغَارِبِ (he threw his rope on the withers) was rendered idiomatically as *he let things slide*, whereas the baseline output produced *he let things go uncontrolled*, preserving the propositional meaning but reducing figurative richness (i.e., metaphorical imagery and expressive force were neutralised).

These examples show that ChatGPT can translate opaque idioms effectively, although the richness of the figurative imagery depends on whether the prompt explicitly signals that the expression is an idiom.

In other cases, translations became explanatory when no idiom-aware hint was given. For instance, أَكَلَ بِعَقْلِهِ حَلَاوَةَ (he ate sweets with his mind) was rendered in the baseline condition as *he easily deceived him*, which conveys the intended meaning but loses the idiomatic imagery. Similarly, بَيْتَ الْقَصِيدِ (the line of the poem) was rendered in the baseline output as *the heart of the matter*, while the idiom-aware output produced *the crux of the matter*, maintaining a stronger figurative detail. A comparable example appears in وَقَعَ فِي حَيْصِ بَيْصِ (he fell into hay bays). With the idiom-aware prompt, ChatGPT produced *he was in a pickle*, preserving both the figurative meaning and the idiomatic form, whereas the baseline output provided a more explanatory rendering — *he was in great confusion*. Likewise, أَلْقَى الْكَلَامَ عَلَى عَوَاهِيهِ (he threw out words on their flaws) was rendered in the idiom-aware condition as *spoke without due care or consideration*, whereas

the baseline output produced *he spoke recklessly*, giving an explanatory rendering with reduced figurative imagery. These cases indicate that ChatGPT tends to rely on explanatory translations when rendering opaque idioms unless the prompt explicitly signals that the expression is idiomatic. In addition, a few expressions preserved the intended meaning but lacked idiomatic form in both conditions. For example, رَفَعَ عَقْبِرَتَهُ (raised his voice) was rendered as *raised his voice* in both baseline and idiom-aware prompts. Although the propositional meaning is conveyed, the figurative sense is not retained.

Several other opaque idioms show similar patterns when baseline and idiom-aware outputs are compared. The idiom عَادَ بِخُفَّيْ حُنَيْنٍ (he returned with Hunayn's two sandals) was rendered idiomatically as *returned empty-handed*, while the baseline output produced *he came back empty-handed*, which conveys the literal meaning but weakens the figurative semantics.

In some cases, ChatGPT produced the same result in both conditions. The idiom اَنْزَلَ مِنْ طِينٍ وَاَنْزَلَ مِنْ عَجِينٍ (an ear of clay and an ear of dough) was rendered in both baseline and idiom-aware outputs as *he turned a deaf ear*, and نَسِيبٌ وَحْدَهُ (a weave of his own) was rendered in both cases as *one of a kind*. Opaque idioms proved more challenging for ChatGPT, since their figurative meanings are not inferable from the literal imagery. As a result, the system relied more on descriptive or explanatory phrasing when translating Arabic idioms under the baseline prompt.

The examples support the patterns identified earlier: baseline translations rely on literal or explanatory renderings that preserve propositional meaning but reduce figurative richness (i.e., metaphorical imagery and expressive force are neutralised). By contrast, idiom-aware prompts allow ChatGPT to produce more idiomatic English expressions that retain both meaning and figurative force.

These cases demonstrate that idiom substitution must be assessed for conceptual equivalence. An English idiom that appears structurally parallel may still alter the semantic focus of the source expression. Consequently, translation cannot be equated with the mere presence of an idiom in the target language. A distinction must therefore be maintained between semantic equivalence, idiomatic naturalness, and retention of metaphorical logic. Where these criteria diverge, the translation should be treated as a partial match rather than a fully successful rendering.

While earlier research has shown that neural systems tend to translate transparent idioms more successfully than opaque ones (Dankers et al., 2022; Fadaee et al., 2018), the present study extends this line of research in three ways. First, the findings current knowledge of prompt sensitivity in large language models (LLMs). Previous studies demonstrate that prompt design affects translation quality (Castaldo & Monti, 2024; Zhang et al., 2023), but they do not systematically distinguish between idiom types within

‘Human oversight appears most necessary in three types of cases identified in the corpus: (i) idioms involving culturally specific historical or narrative references, where cultural imagery is systematically neutralised; (ii) idioms encoding causal or evaluative nuance that may be semantically shifted through substitution; and (iii) idioms whose pragmatic force depends on register or expressive intensity, where paraphrase weakens rhetorical impact. In these instances, semantic adequacy does not guarantee figurative fidelity, and human review is required to ensure conceptual alignment.’

a culturally distant language pair such as Arabic–English. This study shows that prompt effects are not uniform: idiom-aware prompting yields only marginal improvements for transparent idioms but produces richer figurative equivalents for opaque idioms. In other words, prompt sensitivity interacts with the degree of semantic transparency. This interaction has not yet been explicitly theorised in previous research on idiom translation in LLMs.

Second, the study contributes a culturally specific perspective to idiom translation research. Much existing LLM research focuses on European language pairs or other high-resource settings. Arabic idioms, however, are often deeply entrenched in religious, historical, and Bedouin imagery. The analysis shows that ChatGPT systematically prioritises communicative adequacy in English over the preservation of culturally specific imagery. This reveals an important tendency: LLMs optimise for target-language naturalness rather than the retention of cross-cultural metaphorical imagery. This cultural-flattening tendency, observed consistently across opaque idioms, advances discussions of ‘cultural filtering’ (House, 2015) in AI-mediated translation and demonstrates that figurative fidelity and communicative fluency can operate in tension.

Third, the study proposes a qualitative strategy—error framework tailored to LLM idiom evaluation. Rather than relying on automatic metrics such as BLEU, the analysis integrates translation strategy coding (Baker, 2018) with error typology (literalism, calque, semantic mismatch) and assessments of cultural adequacy. This framework provides a replicable model for evaluating idioms in LLM translation, particularly for low-frequency and culturally dense expressions. The study thus offers a structured evaluative approach instead of purely descriptive observation.

Thus, the contribution of this study does not lie simply in showing that transparent idioms are easier to translate or that prompting improves idiomaticity — both

findings are acknowledged in earlier research. Rather, the analytical advance lies in demonstrating (i) the interaction between prompt sensitivity and idiom transparency, (ii) the systematic tendency of LLMs to replace culturally situated imagery with conventional target-language idioms, and (iii) a qualitative evaluation model suitable for culturally bound figurative language. The corresponding findings clarify how LLMs handle non-compositional expressions and identify the conditions under which human oversight remains necessary.

At the same time, some limitations should be acknowledged. Given the limited scope of the dataset and the model setup, the generalisability of the findings is necessarily constrained; accordingly, the conclusions should be interpreted as indicative patterns within a controlled pilot corpus rather than as definitive claims about LLM idiom translation more broadly. While ChatGPT shows notable capacity to render transparent idioms and, with prompting, even some opaque idioms, figurative and cultural fidelity remains uneven. These observations have implications for translation pedagogy, applied linguistics, and the design of future MT systems that account more effectively for cultural and figurative aspects of language.

6. CONCLUSION

This study investigated how ChatGPT translates Arabic idioms into English, focusing on the strategies it employs, the errors it makes, and its accuracy across transparent and opaque idioms. Using Moon’s (1998) classification, 26 idioms were collected and translated under two conditions: a baseline prompt and an idiom-aware prompt, not to measure output against a human translator, but to critique the system’s own renderings systematically. Beyond descriptive findings, the dataset allows more precise pedagogical and professional implications to be drawn.

Human oversight appears most necessary in three types of cases identified in the corpus: (i) idioms involving culturally specific historical or narrative references, where cultural imagery is systematically neutralised; (ii) idioms encoding causal or evaluative nuance that may be semantically shifted through substitution; and (iii) idioms whose pragmatic force depends on register or expressive intensity, where paraphrase weakens rhetorical impact. In these instances, semantic adequacy does not guarantee figurative fidelity, and human review is required to ensure conceptual alignment.

Study findings also indicate that prompt design affects translation reliability. Idiom-aware prompting consistently reduced literalism and increased the likelihood of established English equivalents, particularly for opaque idioms. However, even idiom-aware prompts did not preserve source imagery; instead, they favoured culturally familiar English substitutions. This suggests that a minimally effective prompting strategy for practitioners is to

explicitly signal figurative intent (e.g., ‘Translate the following idiom into natural English’), while recognising that such prompts optimise target-language idiomaticity rather than the retention of source-culture metaphorical imagery.

For translation pedagogy, these findings imply that LLM outputs should be treated as candidate drafts rather than final products, especially when working with culturally embedded expressions. Transparent idioms with conventional English parallels may be used with minimal revision, whereas opaque idioms, metaphorically dense constructions, and historically rooted expressions require careful human evaluation. The pedagogical value of AI, therefore, lies in accelerating the generation of plausible

equivalents that still require critical assessment for semantic precision and figurative richness. Ultimately, this study advances idiom translation research by demonstrating that prompt effects are conditioned by semantic transparency and that LLMs systematically privilege target-language naturalness over source-culture imagery. While ChatGPT performs strongly at the level of communicative adequacy, it frequently neutralises culturally embedded metaphor. This suggests that future LLM development must address not only idiomatic correctness but also the retention of figurative and cultural content. AI can assist in idiom translation, but culturally faithful transfer remains a task that requires human mediation.

References

- Adelnia, A., & Dastjerdi, H. V. (2011). Translation of idioms: A hard task for the translator. *Theory and Practice in Language Studies*, 1(7), 879–883. <https://dx.doi.org/10.4304/tpls.1.7.879-883>
- Al-Maydānī, A. M. (2005). *Majmā al-amthāl*. Dār Ṣādir.
- Al-Zabīdī, M. M. (2011). *Tāj al-ʿarūs min jawāhīr al-qāmūs*. Dār Ṣādir.
- Al-ʿAskarī, A. H. (1964). *Jamharat al-Amthāl*. Dār al-Maʿārif.
- 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://dx.doi.org/10.36892/ijlls.v7i3.2151>
- Almaaytah, S. A. (2022). Translation of idiomatic expressions from Arabic into English using AI (Artificial Intelligence). *Journal of Positive School Psychology*, 6(4), 8839–8846.
- Almahasees, Z. (2021). *Analysing English-Arabic machine translation: Google Translate, Microsoft Translator and Sakhr*. Routledge. <https://doi.org/10.4324/9781003191018>
- Arabic Language Academy. (2004). *Al-Mūjam al-wasīṭ* (4th ed.). Maktabat al-Shurūq al-Dawliyya.
- Baker, M. (1992). *In other words: A coursebook on translation*. Routledge. <https://doi.org/10.4324/9780203133590>
- Baker, M. (2018). *In other words: A coursebook on translation* (3rd ed.). Routledge. <https://dx.doi.org/10.4324/9781315619187>
- Baziotis, C., Mathur, P., & Hasler, E. (2022). Automatic evaluation and analysis of idioms in neural machine translation. *arXiv*, 2210.04545. <https://dx.doi.org/10.48550/arXiv.2210.04545>
- Castaldo, A., & Monti, J. (2024, June). Prompting large language models for idiomatic translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology* (pp. 37–39). European Association for Machine Translation.
- CD. (2026a). Be in full swing. *Cambridge Dictionary*. <https://dictionary.cambridge.org/dictionary/english-russian/be-in-full-swing>
- CD. (2026b). Go up in smoke. *Cambridge Dictionary*. <https://dictionary.cambridge.org/dictionary/english/go-up-in-smoke>
- CD. (2026c). Lose your mind. *Cambridge Dictionary*. <https://dictionary.cambridge.org/dictionary/english/lose-mind>
- Dankers, V., Lucas, C. G., & Titov, I. (2022). Can transformer be too compositional? Analysing idiom processing in neural machine translation. *arXiv*, 2205.15301. <https://doi.org/10.48550/arXiv.2205.15301>
- Darwish, N., Haider, A., Tannous, B., Rumman, R. N. A., Alantari, D., Saed, H., & Dagamseh, M. (2025). A reception study of AI-translated idioms and proverbs between Arabic and English. *Research Journal in Advanced Humanities*, 6(3). <https://dx.doi.org/10.58256/k4d6pp20>
- Fadaee, M., Bisazza, A., & Monz, C. (2018). Examining the tip of the iceberg: A data set for idiom translation. *arXiv*, 1802.04681. <https://dx.doi.org/10.48550/arXiv.1802.04681>
- Ghazala, H. (2008). *Translation as problems and solutions: A textbook for university students and trainee translators*. Dar El-Ilm Lilmalayin.
- Hatim, B., & Mason, I. (1990). *Discourse and the translator*. Longman. <https://doi.org/10.4324/9781315846583>
- Hinds, M., & Badawi, E.-S. (1986). *A dictionary of Egyptian Arabic: Arabic–English*. Librairie du Liban.
- House, J. (2015). *Translation quality assessment: Past and present*. Routledge. <https://doi.org/10.4324/9781315752839>
- Ibn Manzūr, M. M. (1956). *Lisān al-ʿArab*. Dār Ṣādir.
- Liu, E., Chaudhary, A., & Neubig, G. (2023, December). Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 15095–15111). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.933>
- Makhmudova, Z., & Khamitov, E. (2025). Idioms, proverbs, and metaphors: Cultural reflections in language. *Modern American Journal of Linguistics, Education, and Pedagogy*, 1(2), 180–187.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*. SAGE.
- Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.
- Newmark, P. (1988). *A textbook of translation*. Prentice Hall.

- Nida, E. A. (1964). *Toward a science of translating: With special reference to principles and procedures involved in Bible translating*. Brill.
- Ramisch, C. (2015). *Multiword expressions acquisition: A generic and open framework*. Springer. <https://dx.doi.org/10.1007/978-3-319-09207-2>
- Sadigzade, Z. (2025). Idiomatic expressions and their impact on lexical competence. *Journal of Azerbaijan Language and Education Studies*, 2(1), 26–33.
- Saldanha, G., & O'Brien, S. (2014). *Research methodologies in translation studies*. Routledge.
- Sandelowski, M. (2000). Whatever happened to qualitative description? *Research in Nursing & Health*, 23(4), 334–340.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.
- Yamada, M. (2023). Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. *arXiv*, 2308.01391.
- Yousef, N. T. (2024). Strategies used in Arabic-English translation of idioms in Samiha Kraisa's novel *Al Qurmiya*. *World Journal of English Language*, 14(2), Article 271. <https://doi.org/10.5430/wjel.v14n2p271>
- Zaninello, A., & Birch, A. (2020, May 11–16). Multiword expression-aware neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (pp. 3816–3825). European Language Resources Association.
- Zhang, B., Haddow, B., & Birch, A. (2023, July). Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 41092–41110). PMLR.
- Zhu, Y., Lal, D. M., Denysiuk, S., & Mitkov, R. (2024, July 3-6). From neural machine translation to large language models: Analysing translation quality of Chinese idioms. In *Proceedings of New Trends in Translation and Technology* (pp. 247–260). BACL. https://dx.doi.org/10.26615/issn.2815-4711.2024_021

ABOUT THE AUTHORS

Mohd Nour Al Salem

PhD in Translation, Associate Professor of Applied Linguistics
Department of English Language and Literature, School of Foreign Languages
The University of Jordan, Amman, Jordan
Postal address: 21942 Amman, Jordan
Email: mo.alsalem@ju.edu.jo
ORCID ID: <https://orcid.org/0000-0002-7007-6444>

Nimer Abusalim

PhD in Linguistics, Professor of Linguistics
Department of English Language and Literature, School of Foreign Languages
The University of Jordan, Amman, Jordan
Postal address: 21942 Amman, Jordan
Email: n.abusalim@ju.edu.jo
ORCID ID: <https://orcid.org/0000-0003-2733-7693>

Sharif Alghazo

PhD in Applied Linguistics, Professor of Linguistics
Department of Foreign Languages, College of Arts, Humanities, and Social Sciences, University of Sharjah, Sharjah, UAE
Department of English Language and Literature, School of Foreign Languages, The University of Jordan, Amman, Jordan
Postal address: F5 University Street, University of Sharjah, UAE
Email: [salghazo@sharjah.ac.ae](mailto:salg hazo@sharjah.ac.ae)
ORCID ID: <https://orcid.org/0000-0002-8163-283X>