

RESEARCH ARTICLE

Evaluating GPT-5 Literary Metaphor Translation from English into Arabic through Recognition, Interpretation, and Rendering

Hani Hamad M Albelihi¹ and Arwa Alsaleh²

¹ Department of English Language and Literature, College of Languages and Humanities, Qassim University, Saudi Arabia
<https://orcid.org/0000-0001-5186-9687>

² Department of English Language and Literature, College of Languages and Humanities, Qassim University, Saudi Arabia
<https://orcid.org/0000-0003-3994-2196>

Corresponding Author: Hani Hamad M Albelihi, **E-mail:** h.albelihi@qu.edu.sa

ABSTRACT

Research has indicated that artificial intelligence (AI) systems can produce fluent literary translations but may struggle to preserve metaphorical meaning and imagery. However, most previous studies have evaluated completed translations, providing limited insight into whether AI systems recognize and interpret metaphors more successfully than they render them. This study extends previous research by examining GPT-5's recognition, interpretation, and Arabic rendering of 202 literary metaphors identified through MIP/MIPVU in the first two chapters of Kazuo Ishiguro's *The Buried Giant*. GPT-5's Arabic renderings were also compared with those produced by DeepL and Kholoud Amr's published translation. Each dimension was evaluated using a 0–3 ordinal scale, with high inter-rater reliability (weighted Cohen's $\kappa = .89-.93$). The results showed that GPT-5 performed better in metaphor recognition ($M = 2.30$) and interpretation ($M = 2.29$) than in rendering ($M = 1.37$). DeepL achieved higher rendering scores ($M = 1.68$), while the published translation performed substantially better ($M = 2.55$). Qualitative analysis identified recurring patterns involving lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation. The findings suggest that successful metaphor recognition and interpretation do not necessarily correspond to effective figurative rendering in Arabic, highlighting the value of evaluating these dimensions separately.

KEYWORDS

Literary metaphor translation; generative artificial intelligence; GPT-5; translation quality assessment; figurative language; English–Arabic translation.

ARTICLE INFORMATION

ACCEPTED: 24 September 2026

PUBLISHED: 24 October 2026

DOI: 10.32996/ijllt.2026.9.10.26

1. Introduction

Research on artificial intelligence (AI) in literary translation has developed rapidly, with neural machine translation (NMT) systems and large language models (LLMs) producing fluent output (Gao et al., 2024; Guerberof-Arenas & Toral, 2022; Zhang et al., 2025). Literary metaphor remains challenging, because figurative expressions convey meanings extending beyond their literal forms (Newmark, 1988; Schäffner, 2004; van den Broeck, 1981). AI-generated translations may preserve general meaning while weakening imagery, cultural associations, and rhetorical effects (Karakanta et al., 2025; Manapbayeva et al., 2024; Zhang et al., 2025).

Metaphor translation quality is multidimensional, involving conceptual meaning, imagery, and rhetorical effect (Newmark, 1988; Schäffner, 2004; Wang et al., 2024). Three dimensions are particularly relevant: recognition, identifying an expression as metaphorical; interpretation, explaining its meaning in narrative context; and rendering, expressing that meaning in the target language while preserving its imagery, tone, and rhetorical effect (Kövecses, 2010; Lakoff & Johnson, 1980; Mandelblit, 1995). These do not necessarily represent independent stages of AI processing: a system may recognize and explain a metaphor accurately without producing an equally effective translation. Examining them separately may therefore yield a more detailed assessment than the completed translation alone. Existing research falls into three strands. The first addresses the limitations of conventional machine translation evaluation: Freitag et al. (2021) showed that expert evaluation identifies meaning-related

problems that automatic methods overlook, and Wang et al. (2024) developed the MMTE framework to assess metaphorical equivalence, emotion, authenticity, and overall quality.

The second strand has compared LLMs, NMT systems, and professional translators, with mixed results. Gao et al. (2024) found ChatGPT performed better than Google Translate and DeepL on Chinese classical poetry, whereas Manapbayeva et al. (2024) found that a professional translation of Abai's "Spring" conveyed cultural and poetic nuance more effectively than ChatGPT-4. Karakanta et al. (2025) found divergence between automatic metrics and expert judgment in English–Dutch metaphor translation, DeepL scoring slightly above GPT-4, while Zhang et al. (2025) reported advantages for published translations. Performance thus appears to depend on language pair, genre, prompting, and evaluation criteria.

The third strand has examined AI translation of figurative and pragmatic meaning involving Arabic. Al Rousan et al. (2025) found ChatGPT fluent but weaker on culturally dependent meanings; Sahari et al. (2024) reported human advantages in idiomaticity and cultural sensitivity; Sharar et al. (2025) identified difficulties with connotative meaning in Arabic literary prose; and Shehab and Badah (2026) found AI renderings of conversational implicatures preserved surface meaning while weakening pragmatic force.

Across all three strands, evaluation has centered on the completed translation, and little attention has been given to whether successful identification and explanation of metaphors correspond to successful rendering. The English–Arabic pair is particularly relevant because the two languages differ in linguistic structure, rhetorical convention, and figurative tradition (Abdul-Raof, 2006; Samaniego Fernández, 2022; Schäffner, 2004), creating difficulties with lexical ambiguity, personification, and culturally specific imagery. Metaphors may also recur across a narrative, making their translation depend on more than the immediate sentence.

This study uses a stage-aware framework to evaluate 202 metaphorical expressions identified through MIP/MIPVU in the first two chapters of Kazuo Ishiguro's *The Buried Giant* (2015). GPT-5's recognition, contextual interpretation, and Arabic rendering are scored on separate 0–3 ordinal scales, and its rendering is compared with DeepL and Kholoud Amr's published translation (Ishiguro, 2018). The framework examines differences among observable outputs rather than assuming independent cognitive stages. The study addresses three research questions:

RQ1. To what extent does GPT-5 recognize, interpret, and render the literary metaphors in *The Buried Giant*, and what differences are observed across these dimensions?

RQ2. How do the Arabic metaphor renderings produced by GPT-5 and DeepL compare with each other and with Amr's published translation in rendering quality and failure types?

RQ3. What recurrent linguistic and translational patterns characterize successful and unsuccessful metaphor rendering in GPT-5 and DeepL?

2. Literature review

2.1 Metaphor translation

Earlier approaches to metaphor translation focused on translators' available procedures: preserving the source image, replacing it with another metaphor, resorting to literal paraphrase, or omitting the figure altogether (Newmark, 1988; Schäffner, 2004; van den Broeck, 1981). Cognitive approaches later reframed metaphor as a conceptual mapping that varies across languages and cultures, shifting the analytic focus from procedure to process (Kövecses, 2010; Lakoff & Johnson, 1980; Schäffner, 2004). This process-oriented turn provides a basis for treating recognition, interpretation, and rendering as separate analytical dimensions in the present study (Mandelblit, 1995; PACTE, 2003). The literature on human translation indicates that comprehension and production rely on distinct skills: a translator might understand a figurative meaning but still produce a literal or flat rendering, so output quality cannot necessarily be read back as evidence of understanding. Mandelblit's Cognitive Translation Hypothesis extends this point, arguing that when source- and target-language conceptual mappings diverge, translation demands more effort and tends toward literalness (Mandelblit, 1995).

2.2 Metaphor loss and rendering outcomes

Metaphorical meaning relies on lexical form, convention, culture, co-text, and context. Thus, a translation may convey the overall situation without preserving the original image, allowing metaphorical loss to remain hidden in otherwise fluent and plausible output (Abdul-Raof, 2006; Samaniego Fernández, 2022; Schäffner, 2004). Two recurrent forms of such loss are de-metaphorization and conceptual substitution. De-metaphorization omits the metaphor outright, reducing imagery and emotional impact; conceptual substitution retains a figurative expression but shifts it to a different image, altering the figurative meaning while sometimes preserving vividness (Newmark, 1988; Samaniego Fernández, 2022; Schäffner, 2004). Such changes matter most when the metaphor recurs as part of a narrative pattern.

Building on this distinction, we categorize outcomes as recognition, interpretation, and rendering rather than simply pass/fail: preservation does not guarantee accuracy or naturalness, and substitution is not inherently a failure if it is effective. Rendering type and translation quality are therefore treated as related but distinct: the former describes how the source mapping changes, while the latter assesses accuracy and naturalness.

2.3 Toward metaphor-sensitive evaluation

Traditional machine-translation metrics are of limited use for evaluating literary metaphors because they focus mainly on word overlap or overall semantic accuracy. Such metrics may reward translations that stay close to source wording while losing imagery, voice, or figurative force, and may penalize effective literary reformulations that depart from a reference translation's exact wording (Freitag et al., 2021; Karakanta et al., 2025; Wang et al., 2024). Recent evaluation research has moved beyond fluency and propositional accuracy toward human, context-sensitive, and figuratively informed assessment.

Wang et al.'s (2024) MMTE framework combines a multilingual metaphor corpus with human evaluation across four dimensions: metaphorical equivalence, emotion, authenticity, and overall quality, because mainstream metrics pay limited attention to figurative quality. Freitag et al. (2021) similarly demonstrate that expert human evaluation surfaces contextual and meaning-related problems that automatic, reference-based measures overlook. Extending this line of inquiry to literary metaphor specifically, Karakanta et al. (2025) conducted a mixed-methods evaluation of several NMT systems and LLMs translating English literary metaphors into Dutch from the VU Amsterdam Metaphor Corpus, reporting divergence between automatic metrics and expert judgment in system ranking: human evaluators rated a literary-adapted system at 75% accuracy, DeepL at 74%, and GPT-4 at 70%.

These studies establish the value of human, metaphor-sensitive evaluation, but they assess completed translations and therefore cannot determine whether an unsuccessful rendering is due to recognition, interpretation, or target-language realization (Freitag et al., 2021; Karakanta et al., 2025; Wang et al., 2024). The present study extends this work by evaluating these three outputs separately for GPT-5 and comparing final renderings with DeepL and the published translation (Karakanta et al., 2025; Mandelblit, 1995).

2.4 Variation across LLM, NMT, and human translation

No consensus yet exists on which AI system is superior for literary translation across contexts. Context window, prompting strategy, genre, model architecture, and language pair all shape whether the target text preserves, paraphrases, replaces, literalizes, or omits figurative language. For instance, Gao et al. (2024) found that ChatGPT's translations of Chinese classical poetry improved when prompts explicitly addressed rhythm, rhyme, and style. In the same vein, Manapbayeva et al. (2024) reported that although ChatGPT-4 produced a coherent translation of Abai's poem "Spring," the professional human translation better conveyed its cultural, emotional, and poetic nuance.

Metaphor translation also varies systematically across languages. Translating from English into Arabic raises difficulties with metaphors built on lexical ambiguity, cultural imagery, personification, or extended mappings that lack direct equivalents (Abdul-Raof, 2006; Samaniego Fernández, 2022; Schäffner, 2004). Evidence across several language pairs points to advantages for published human translation in idiomaticity, cultural nuance, and lexical richness, while model output tends to remain more literal or structurally closer to the source (Al Rousan et al., 2025; Manapbayeva et al., 2024; Sahari et al., 2024; Zhang et al., 2025). Within this pair, Al Rousan et al. (2025) offered one of the closest precedents, comparing ChatGPT with a published human translation of a modern Arabic novel using an adapted Multidimensional Quality Metrics scheme. They found ChatGPT fluent but consistently less accurate than the professional translation, especially where meaning depended on cultural and pragmatic inference. That study, however, translates from Arabic into English, evaluates only the finished text, and includes no dedicated NMT comparator. Thus, the English-to-Arabic direction, the relationships among separately elicited recognition, interpretation, and rendering outputs, and the behavior of a specialized NMT system remain largely unexplored.

Related work on Arabic figurative and pragmatic language points in the same direction. Sahari et al. (2024) found human translators of figurative Arabic–English material more idiomatic and culturally sensitive than ChatGPT; Sharar et al. (2025) found many of ChatGPT's renderings of culturally embedded expressions in Arabic literary prose inadequate because the model leaned on denotative meaning at the expense of connotation; and Shehab and Badah (2026) similarly reported that AI translations of Arabic literary implicatures often preserve surface meaning while failing to reproduce the pragmatic force that professional translators convey more successfully.

These findings should not be read as evidence that human superiority is automatic or independent of research design; rather, conclusions depend on evaluator expertise, textual context, and the granularity of the evaluation framework. Three limitations remain particularly relevant to the present study. First, most metaphor-sensitive frameworks assess completed translations rather than separately elicited recognition, interpretation, and rendering outputs. Second, the discrepancy between successfully interpreting figurative meaning and successfully rendering it in the target language remains underexamined in generative AI translation. Third, English–Arabic literary metaphor remains underexplored in studies comparing a general-purpose LLM, a dedicated NMT system, and a published professional translation within the same corpus (Al Rousan et al., 2025; Freitag et al., 2021; Karakanta et al., 2025; Mandelblit, 1995; Wang et al., 2024).

2.5 Present study

To address these limitations, we evaluate 202 literary metaphors from Ishiguro's *The Buried Giant* (Ishiguro, 2015) at two levels: GPT-5 recognition and interpretation, and comparative rendering across GPT-5, DeepL, and Amr's published translation

(Ishiguro, 2018). We assess recognition and interpretation only for GPT-5 because we cannot directly elicit them from DeepL or reliably reconstruct them from the published translation.

The study addresses the following research questions:

RQ1. To what extent does GPT-5 recognize, interpret, and render the literary metaphors in *The Buried Giant*, and where are the largest differences observed across these analytical dimensions?

RQ2. How do the Arabic metaphor renderings produced by GPT-5 and DeepL compare with each other and with Kholoud Amr's published Arabic translation in terms of rendering quality and failure types?

RQ3. What recurrent linguistic and translational factors characterize successful and unsuccessful metaphor rendering in GPT-5 and DeepL?

3. Methodology

This study employed a mixed-methods design to evaluate GPT-5's recognition, interpretation, and Arabic rendering of literary metaphors and to compare its rendering performance with that of DeepL and a published professional translation. Quantitative analysis examined differences among GPT-5's three elicited output dimensions and compared rendering quality across the three translation sources. Qualitative textual analysis investigated recurrent linguistic and translational patterns associated with successful and unsuccessful metaphor rendering (Creswell & Plano Clark, 2018).

The study used a stage-aware analytical framework, evaluating recognition, interpretation, and rendering as distinct aspects of observable output quality. It did not assume that these dimensions reflected independent cognitive processes or experimentally isolated stages of translation.

3.1 Corpus and translation sources

Kazuo Ishiguro's *The Buried Giant* (2015) was selected because of its extensive use of metaphor, symbolism, allegory, and imagery, making it suitable for investigating figurative language in literary translation (Borowska-Szerszun, 2016; Karakanta et al., 2025; Li & Li, 2021). The first two chapters provided a bounded literary corpus within which contextual meaning and recurring imagery could be examined. The original identification and adjudication procedure yielded a dataset of 202 expressions classified as metaphorical.

Three translation sources were examined. GPT-5 generated metaphor identifications, contextual interpretations, and Arabic translations. DeepL provided English-to-Arabic machine translations, while Kholoud Amr's published Arabic translation of *The Buried Giant* (Ishiguro, 2018) served as a professional literary comparator. The published translation, which received third prize in the English-into-Arabic category of the 2019 Sheikh Hamad Award for Translation and International Understanding, provided an established example of professional translation practice.

The published translation was not treated as an error-free reference or definitive gold standard. Instead, all three sources were evaluated using the same rendering-quality rubric. Because their production conditions differed, the comparison was designed to examine the quality of the resulting Arabic renderings rather than to establish the intrinsic superiority of a particular translation technology or translation process.

3.2 Metaphor identification

We identified metaphorical expressions using the Metaphor Identification Procedure (MIP; Pragglejaz Group, 2007), supplemented by principles from MIPVU (Steen et al., 2010). The identification procedure involved examining lexical units in their narrative context, determining their contextual meanings, establishing more basic meanings, and assessing whether the contextual meanings contrasted with the basic meanings while remaining understandable through comparison.

Particular attention was given to distinguishing linguistic metaphor from literary symbolism, allegory, and literal reference. An expression was not considered metaphorical solely because it contributed to the novel's symbolic or thematic structure. Instead, metaphorical status depended on evidence of a comparison-based relationship between the lexical unit's contextual and basic meanings.

Identification was conducted at the lexical-unit level, while the complete expression and relevant narrative context were retained for subsequent translation analysis. This distinction was necessary because the metaphorical meaning of an individual lexical unit may depend on a broader expression, and its translation may involve grammatical, lexical, or rhetorical restructuring in Arabic.

Two researchers independently examined the source text. Initial agreement was reported as 92%, and disagreements were resolved through discussion. The resulting dataset contained 202 expressions classified as metaphorical under the original procedure.

For the purposes of construct validity, potentially ambiguous expressions, particularly those involving symbolic objects, fictional entities, or recurring character references, require verification against the MIP/MIPVU criteria. This distinction is especially important when the literary significance of an expression extends beyond its linguistic meaning. The classification records should document contextual meaning, basic meaning, comparison-based relationships, and the justification for including borderline cases.

A supplementary identification audit should document these decisions and establish whether the classifications retained in the analytical corpus satisfy the stated criteria. Any resulting changes to the corpus would require corresponding revisions to the reported sample size and affected analyses.

3.3 Data collection

GPT-5 was accessed through the OpenAI ChatGPT web interface on October 13, 2025 (OpenAI, 2025). The interface did not expose model-generation parameters, including temperature, so these settings could not be independently controlled or recorded.

For each target passage, the source text was submitted in a fresh conversation. This procedure provided narrative context for interpreting the target expressions while minimizing contextual carryover from previously analyzed passages. A single response was collected for each passage, without regeneration or manual selection among alternative outputs. Consequently, the findings represent performance under the recorded prompting conditions rather than estimates of variability across repeated generations.

A standardized prompt was used to elicit three analytically distinct outputs: metaphor recognition, contextual interpretation, and Arabic rendering:

Act as a professional literary translator. Identify the metaphorical expressions in the passage below, explain their meanings within the context of the story, and translate the passage into natural literary Arabic while preserving, as far as possible, its metaphors, imagery, tone, and figurative effects.

The standardized prompt elicited recognition, interpretation, and rendering within a single response. These outputs were evaluated as analytically distinct dimensions of observable model performance rather than as independent experimental conditions or sequential cognitive stages. Because the tasks were elicited together, the resulting outputs may reflect dependencies introduced by the prompting procedure. Accordingly, differences among recognition, interpretation, and rendering were interpreted as performance discrepancies within the specified prompting condition, rather than as evidence of independent processing mechanisms or causal relationships among the three dimensions.

GPT-5's responses were retained without manual correction. DeepL translations were produced on the same day using the English-to-Arabic setting (DeepL SE, 2025). Recognition and interpretation were assessed only for GPT-5 because equivalent outputs were not elicited from DeepL and could not be reconstructed reliably from the published translation. Rendering quality was evaluated for all three sources.

The comparison was therefore asymmetric by design: GPT-5 was assessed across three elicited dimensions, whereas all three translation sources were compared exclusively on final Arabic rendering quality. Furthermore, the sources were not generated under identical production conditions. GPT-5 received explicit literary and metaphor-preservation instructions, while the professional translation was an independently produced published work. Such differences were considered when interpreting cross-source performance.

All reported Arabic examples were preserved as generated or published, without editorial correction.

Each target expression was recorded in an Excel dataset containing its source sentence, relevant narrative context, GPT-5 recognition and interpretation outputs, GPT-5 and DeepL renderings, and the corresponding segment from the published Arabic translation. Translation alignment was based on complete metaphorical expressions rather than mechanical word-level correspondence.

3.4 Evaluation and coding

The evaluation framework assessed recognition, interpretation, and rendering using separate 0–3 ordinal scales (Table 1). Recognition evaluated whether GPT-5 explicitly identified a preidentified target expression as metaphorical (Pragglejaz Group, 2007; Steen et al., 2010). Interpretation evaluated the accuracy and contextual relevance of the explanation provided (Kövecses, 2010; Lakoff & Johnson, 1980; Steen, 1999). Rendering evaluated semantic accuracy, grammaticality, naturalness, imagery, tone, stylistic appropriateness, and preservation of figurative meaning (Karakanta et al., 2025; Lommel et al., 2014; Samaniego Fernández, 2022; Schäffner, 2004; Wang et al., 2024).

Although the three dimensions used the same numerical range, each represented a distinct construct. Recognition concerned identifying the target expression, interpretation concerned explaining it in context, and rendering concerned realizing it in Arabic. The scores therefore represented separate ordinal judgments rather than interchangeable measures of translation performance.

Because recognition was evaluated against preidentified target expressions, the recognition scores measured GPT-5's detection of those expressions rather than unrestricted metaphor-classification accuracy. The design did not systematically evaluate false-positive identification of non-metaphorical expressions.

3.5 Rendering quality and failure classification

We assessed rendering quality holistically rather than through independently weighted numerical subcomponents. Semantic accuracy and preservation of central figurative meaning were necessary conditions for scores of 2 or 3. Naturalness, grammaticality, imagery, tone, and rhetorical effectiveness were used to distinguish adequate from excellent renderings.

A score of 3 indicated an accurate, natural, and rhetorically effective Arabic rendering that preserved or appropriately reconstructed the metaphor's conceptual and literary effect. A score of 2 indicated an accurate and generally natural translation that preserved the central figurative meaning but weakened some imagery, tone, or stylistic nuance.

A score of 1 indicated a weak rendering in which the general contextual meaning remained recoverable but the figurative image or rhetorical effect was substantially diminished. This category also accommodated cases of partial metaphor preservation or limited figurative reconstruction.

A score of 0 indicated a substantive failure in rendering. Two failure types were distinguished. Mistranslation (OT) referred to cases in which the translation distorted the contextual meaning of the source expression. Omission or de-metaphorization (OM) referred to cases in which the relevant figurative meaning or image was absent or neutralized without an effective compensatory rhetorical realization.

The distinction between scores 1 and OM depended on how much of the source metaphor's relevant figurative meaning remained recoverable in Arabic. Partial preservation or weakened figurative realization was distinguished from cases where the target expression failed to communicate the relevant metaphorical content. When evaluating potential OM cases, the raters considered whether another rhetorical construction preserved the metaphor's conceptual meaning despite changes in its linguistic form.

Metaphor preservation did not require lexical or structural identity between source and target expressions. We did not penalize reformulations, including metaphor substitution, simile, or other appropriate rhetorical reconstruction, solely because they departed from the source's surface form, provided they preserved the relevant meaning and literary effect.

Translation strategy and rendering quality were conceptually distinguished. Procedures such as metaphor preservation, substitution, paraphrase, or omission describe transformations of figurative language, whereas the rendering score reflects the assessed quality of the resulting translation. No single transformation strategy automatically determined the quality score.

3.6 Inter-rater reliability

Two researchers independently evaluated the recorded outputs. The coding dataset comprised 202 GPT-5 recognition observations, 202 GPT-5 interpretation observations, and 606 rendering observations, consisting of 202 renderings from each translation source.

Because the scores were ordinal, weighted Cohen's kappa was calculated separately for recognition, interpretation, and rendering (Cohen, 1968). For rendering, the reported reliability analysis pooled observations from GPT-5, DeepL, and the published translation.

The initial percentage agreement was 91% for recognition, 88% for interpretation, and 89% for rendering. The corresponding weighted kappa coefficients were .93, .89, and .91, respectively, indicating strong reported inter-rater agreement. Disagreements were subsequently resolved through discussion.

The reliability coefficients describe agreement under the original coding framework. Any substantive changes to the scoring rules or metaphor classifications would require re-examining the affected observations and, where appropriate, recalculating reliability estimates.

3.7 Data analysis

Data analysis combined descriptive statistics, paired nonparametric comparisons, and qualitative textual examination.

3.8 Analysis of GPT-5 Recognition, Interpretation, and Rendering (RQ1)

Descriptive statistics were calculated for the 202 recorded metaphor occurrences, including frequencies, percentages, means, medians, standard deviations, and proportions achieving adequate or excellent scores (≥ 2).

Because the same expressions were evaluated across three ordinal dimensions, the Friedman test was used to examine overall differences among recognition, interpretation, and rendering scores. Significant overall results were followed by pairwise Wilcoxon signed-rank tests with Holm adjustment for multiple comparisons.

We reported Kendall's W as the effect size for the Friedman test, and expressed pairwise effect sizes as $r = |Z|/\sqrt{N}$. The definition of N for these calculations should be documented consistently with the statistical procedure used.

These analyses examined differences among scores for the same target expressions. They were not intended to establish causal or temporal relationships among the model's elicited outputs.

3.9 Comparative Rendering Analysis (RQ2)

We compared rendering scores for GPT-5, DeepL, and the published translation using descriptive statistics and the Friedman test, followed by Holm-adjusted Wilcoxon signed-rank comparisons. Because the three sources translated the same target expressions, the analysis maintained item-level pairing across translation sources.

The frequencies and percentages of zero-score renderings were also examined. Mistranslation (OT) and omission/de-metaphorization (OM) were reported separately to characterize different forms of rendering failure.

Interpretation of these comparisons was limited to the evaluated translations and their documented production conditions. Differences in final scores were not assumed to arise exclusively from model architecture, translator expertise, or any single production-related factor.

3.10 Qualitative Textual Analysis (RQ3)

Qualitative textual analysis examined recurrent cases of successful and unsuccessful metaphor rendering in their narrative context. The analysis drew on established concepts in metaphor translation, including metaphor preservation, de-metaphorization, and conceptual substitution, while also identifying recurring patterns within the dataset.

Four factors were examined: lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation.

Cases were examined across GPT-5, DeepL, and the published translation to investigate how different Arabic realizations preserved, modified, or weakened figurative meaning. Both higher-scoring and lower-scoring examples were considered to illustrate recurring patterns rather than isolated translation errors.

Repeated expressions were examined in relation to their narrative contexts, allowing the qualitative analysis to consider lexical continuity and the development of imagery across occurrences.

3.11 Analytical scope and limitations

The quantitative analyses treated individual metaphor occurrences as the unit of evaluation. However, repeated occurrences of the same metaphor may share lexical, contextual, or discourse-level characteristics. Consequently, observations involving recurring expressions may not be fully independent, and interpreting statistical results should account for the influence of such clustering.

The study also involved a single literary work, a single translation direction, one recorded GPT-5 output per passage, and three translation sources produced under different conditions. These features restrict generalizability and prevent the findings from being interpreted as comprehensive assessments of the relative capabilities of LLMs, neural machine translation, and professional translators.

The study's principal methodological contribution is therefore a structured, output-level evaluation of metaphor identification, explanation, and Arabic rendering under specified conditions, rather than a reconstruction of the internal mechanisms underlying translation performance.

4. Results

This section presents the findings in relation to the three research questions. First, it examines GPT-5's performance across metaphor recognition, contextual interpretation, and Arabic rendering (RQ1). Second, it compares the quality and failure profiles of metaphor renderings produced by GPT-5, DeepL, and the published Arabic translation (RQ2). Third, it examines recurrent linguistic and translational patterns associated with successful and unsuccessful metaphor rendering (RQ3).

The quantitative results are based on 202 target expressions classified as metaphorical in the original dataset. We evaluated recognition, interpretation, and rendering as separate dimensions of GPT-5's observable outputs rather than as independent cognitive or computational stages. The comparisons therefore characterize performance under the specified prompting conditions without establishing causal relationships among the dimensions.

4.1 GPT-5's Performance Across Recognition, Interpretation, and Rendering (RQ1)

Table 1 presents descriptive statistics for GPT-5's recognition, interpretation, and rendering scores.

Table 1

GPT-5's performance across metaphor recognition, interpretation, and rendering

Dimension	n	M	Mdn	SD	≥2 (%)	0 (%)	OM (%)	OT (%)
Recognition	202	2.30	3	1.18	77.2	18.3	—	—
Interpretation	202	2.29	3	1.19	76.7	18.3	—	—
Rendering	202	1.37	1	1.22	50.0	37.6	22.8	14.9

Note. Scores ranged from 0 to 3. ≥2 indicates adequate or excellent performance. OM represents omission/de-metaphorization, while OT represents mistranslation. Percentages may differ slightly from their totals because of rounding.

Recognition and interpretation received similar mean scores ($M = 2.30$ and $M = 2.29$, respectively). 77.2% of recognition cases and 76.7% of interpretation cases showed adequate or excellent performance. In contrast, rendering received a substantially lower mean score ($M = 1.37$, $Mdn = 1$, $SD = 1.22$), with only 50.0% of renderings classified as adequate or excellent.

Zero scores occurred in 37.6% of GPT-5 renderings ($n = 76$). Of these, 46 cases (22.8% of all renderings) were classified as omission or de-metaphorization (OM), and 30 cases (14.9%) were classified as mistranslation (OT). Thus, both loss of figurative meaning and distortion of contextual meaning contributed to the observed rendering difficulties.

The discrepancy between explicit metaphor interpretation and Arabic rendering is illustrated by the expression a welcome antidote to the noon sun. GPT-5 identified antidote as metaphorical and interpreted it as relief from oppressive heat. However, its Arabic rendering was:

كان الريح عنيفًا، لكنه مرحب به تحت شمس الظهيرة

The translation communicates the general idea that the wind provided welcome relief, but the medicinal image conveyed by antidote is not reproduced. This example illustrates a case in which an adequate explicit interpretation was accompanied by a less effective figurative rendering.

To explore whether weaker rendering was restricted to cases of unsuccessful metaphor recognition, rendering scores were examined for the 156 expressions receiving recognition scores of 2 or 3. Within this subset, the mean rendering score was 1.51, compared with 1.37 in the full dataset, and 56.4% of renderings achieved scores of 2 or 3.

These descriptive findings indicate that inadequate rendering was also observed among expressions successfully identified as metaphorical. However, this subset analysis does not establish the underlying causes of rendering difficulty.

Table 2

Inferential comparisons among GPT-5 recognition, interpretation, and rendering scores

Comparison	Holm-adjusted p	Effect size
Omnibus Friedman: $\chi^2(2, N = 202) = 108.95$	< .001	Kendall's W = .27
Recognition vs. Interpretation	.32	—
Recognition vs. Rendering	< .001	$r = .65$
Interpretation vs. Rendering	< .001	$r = .65$

The Friedman test indicated a statistically significant overall difference among the three dimensions, $\chi^2(2, N = 202) = 108.95$, $p < .001$, Kendall's W = .27.

Post hoc Wilcoxon signed-rank tests with Holm adjustment indicated no statistically significant difference between recognition and interpretation (adjusted $p = .32$). In contrast, rendering scores were significantly lower than recognition and interpretation scores (both adjusted $ps < .001$, reported $r = .65$).

These findings answer RQ1 by identifying the largest performance discrepancies between Arabic rendering and the two explicitly elicited analytical dimensions. The results demonstrate that adequate recognition and interpretation do not invariably coincide with adequate target-language rendering under the prompting conditions examined.

4.2 Comparative Metaphor-Rendering Performance (RQ2)

Table 3 presents descriptive statistics for the rendering scores of GPT-5, DeepL, and the published Arabic translation.

Table 3. Metaphor-rendering performance by translation source

Output source	n	M	Mdn	SD	≥ 2 (%)	0 (%)	OM (%)	OT (%)
GPT-5	202	1.37	1	1.22	50.0	37.6	22.8	14.9
DeepL	202	1.68	2	1.18	56.4	23.3	14.9	8.4
Published translation	202	2.55	3	0.97	87.1	10.9	10.9	0.0

Note. All three sources were evaluated using the same rendering-quality scale. The published translation was treated as a professional comparator rather than an error-free reference.

The published Arabic translation achieved the highest mean rendering score ($M = 2.55$, $Mdn = 3$, $SD = 0.97$), followed by DeepL ($M = 1.68$, $Mdn = 2$, $SD = 1.18$) and GPT-5 ($M = 1.37$, $Mdn = 1$, $SD = 1.22$).

Adequate or excellent renderings accounted for 87.1% of the published translation, compared with 56.4% of DeepL renderings and 50.0% of GPT-5 renderings.

The published translation also received the highest proportion of maximum scores (79.2%), compared with DeepL (34.7%) and GPT-5 (24.3%).

Table 4

Inferential comparisons among the three translation sources

Comparison	Holm-adjusted p	Effect size
Omnibus Friedman: $\chi^2(2, N = 202) = 127.23$	< .001	Kendall's W = .31
GPT-5 vs. DeepL	< .001	r = .34
GPT-5 vs. Published translation	< .001	r = .80
DeepL vs. Published translation	< .001	r = .67

The Friedman test revealed a statistically significant difference among the three translation sources, $\chi^2(2, N = 202) = 127.23$, $p < .001$, Kendall's W = .31.

Holm-adjusted Wilcoxon signed-rank tests indicated that DeepL achieved significantly higher rendering scores than GPT-5 (adjusted $p < .001$, reported $r = .34$). The published translation also received significantly higher scores than GPT-5 (adjusted $p < .001$, $r = .80$) and DeepL (adjusted $p < .001$, $r = .67$).

Although DeepL outperformed GPT-5 in the ordinal score comparisons, the difference in adequate-or-excellent rendering rates was 6.4 percentage points. The larger differences were observed between the published translation and both AI-generated translations.

Rendering Failure Profiles. Zero-score renderings accounted for 37.6% of GPT-5 outputs, 23.3% of DeepL outputs, and 10.9% of the published translations.

For GPT-5, 22.8% of renderings were classified as omission/de-metaphorization (0M) and 14.9% as mistranslation (0T). The corresponding percentages for DeepL were 14.9% and 8.4%.

In the published translation, all reported zero-score cases were classified as omission/de-metaphorization (10.9%), with none assigned to mistranslation.

These descriptive differences indicate that the three translation sources exhibited different distributions of rendering failure. However, the results should be interpreted in relation to the study's operational coding definitions, particularly the distinction between weakened figurative realization and complete metaphor loss.

Illustrative Translation Differences

The expression *grey mane* illustrates variation in rhetorical realization. GPT-5 and DeepL rendered the expression as:

شعرها الرمادي

This preserves the descriptive reference to grey hair but removes the animal imagery associated with mane.

The published translator instead used:

شعرها الأشيب مثل لبدة أسد

This version reconstructs the image through a simile, preserving the animal comparison while producing an Arabic expression with a more explicit visual effect.

The example demonstrates that higher-quality figurative rendering need not reproduce the source metaphor through an identical grammatical construction.

A different pattern emerged in the translation of ogres. DeepL consistently used الغيلان, whereas GPT-5 employed alternatives including العفاريت, الكائنات المتوحشة, and الأوغار.

This variation illustrates differences in lexical consistency, although the metaphorical status of ogres must be distinguished from its possible function as a literal reference to fictional creatures.

Similarly, the rendering of boatman as قال القارب in one DeepL output illustrates a referential translation error, since the vessel is substituted for the person. However, whether the underlying English expression constitutes a linguistic metaphor is a separate identification question and cannot be established solely from the translation error.

The published translation was not uniformly successful. The expression *She's got herself a candle* was rendered as:

حصلت لنفسها على شمعة

The translation reproduces the literal reference to a candle but may not fully communicate its broader symbolic significance. Its classification as metaphor loss nevertheless depends on establishing that the source expression itself meets the operational criteria for linguistic metaphor.

Overall, the findings for RQ2 show that the published translation achieved substantially higher rendering-quality scores than either AI system within the evaluated corpus, while DeepL received moderately higher scores than GPT-5. These results

characterize the particular translation products examined and should not be generalized to all literary works, prompting conditions, or translation directions.

4.3 Recurrent Linguistic and Translational Patterns in Metaphor Rendering (RQ3)

Qualitative analysis examined recurring expressions and translation cases to identify linguistic and translational patterns associated with differences in rendering quality.

Four recurring patterns were identified: lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation.

These patterns emerged from the comparative textual analysis and were not independently manipulated or tested as causal predictors of translation success.

Table 5

Mean scores for five recurring expressions by evaluation dimension and translation source

Expression	n	GPT-5 Rec.	GPT-5 Int.	GPT-5 Ren.	DeepL Ren.	Published Ren.
Mist	14	3.00	3.00	2.14	2.29	3.00
Buried Giant	2	3.00	3.00	2.00	2.50	3.00
Candle	14	2.79	2.79	2.14	1.79	2.86
Warren	10	3.00	3.00	0.00	1.20	2.40
Boatman	29	2.90	2.90	0.69	1.62	3.00

The descriptive patterns in Table 5 show variation across recurring expressions. Higher rendering means were observed for mist and candle, while weaker performance was recorded for warren and boatman, especially in GPT-5's outputs. Because the table aggregates repeated occurrences of particular expressions, these descriptive values should not be interpreted as independent observations of metaphor types.

Lexical Equivalence

The availability of an appropriate Arabic lexical equivalent was associated with more stable rendering in some examples. Direct transfer or conventional Arabic equivalence appeared relatively effective when the target expression permitted a recognizable and contextually suitable formulation.

However, lexical correspondence did not guarantee successful translation. A translation could preserve the apparent descriptive meaning of an expression while weakening its figurative image or rhetorical function.

The grey mane example illustrates this distinction. Rendering the expression as شعرها الرمادي communicated the visible attribute but did not reproduce the animal-based comparison.

Discourse-Level Consistency

Recurring expressions also revealed differences in lexical continuity across the narrative.

In the original dataset, the repeated expression boatman received a mean GPT-5 rendering score of 0.69, compared with 1.62 for DeepL and 3.00 for the published translation. GPT-5 and DeepL used different lexical realizations and referential forms, whereas the published translator consistently employed ملاح.

Although the classification of boatman as a linguistic metaphor requires independent verification, these examples illustrate the importance of discourse-level referential consistency in literary translation.

Variability in translation choices may weaken textual continuity when recurring expressions contribute to character identification, narrative cohesion, or the development of imagery.

Contextual Disambiguation

Some rendering difficulties were associated with insufficient differentiation among possible contextual meanings or referents.

The rendering قال القارب for the boatman said is an example of contextual mistranslation, in which a person is replaced by an inanimate object.

Other cases involved the retention of broad semantic content without reproducing the relevant figurative relationship. For example, GPT-5 rendered my fading days through an expression corresponding to my remaining days. Although both refer to a limited period of time, the latter weakens the image of gradual fading.

Such cases demonstrate that preserving general propositional meaning does not necessarily ensure equivalence in the conceptual and expressive content of a metaphor.

Rhetorical Recreation

The clearest qualitative differences appeared in examples where the published translator reconstructed imagery through an alternative Arabic rhetorical form.

In the mask sequence, the published translator used *صخرة صماء* and then developed the imagery with *تفلقت الصخرة وتدفقت منها العواطف*. These expressions produce a coherent figurative progression from emotional rigidity to emotional release.

By comparison, GPT-5 conveyed the emotional meaning more explicitly without reproducing the same progression of imagery.

Conceptual substitution was also observed in the translation of our love withers as *يموت حينا*. Both expressions convey deterioration, but withering emphasizes gradual decline through plant-based imagery, whereas dying invokes a different conceptual endpoint.

Such transformations are not necessarily unsuccessful in every context. Their quality depends on whether the reconstructed figurative expression preserves the relevant conceptual meaning, rhetorical intensity, and literary function.

5. Discussion

This study examined GPT-5's recognition, interpretation, and Arabic rendering of literary metaphors and compared its performance with DeepL and a published professional translation of Kazuo Ishiguro's *The Buried Giant*. The findings indicate three main patterns. First, GPT-5 scored substantially higher on explicitly recognizing and interpreting the target expressions than on rendering them into Arabic. Second, DeepL achieved moderately higher rendering scores than GPT-5, whereas the published professional translation performed substantially better than both AI systems within the evaluated corpus. Third, qualitative analysis identified four recurrent linguistic and translational patterns associated with rendering quality: lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation.

Together, these findings contribute to the evaluation of AI-assisted literary translation by demonstrating the importance of distinguishing explicit metaphor identification and explanation from the quality of figurative realization in the target language. However, because recognition, interpretation, and rendering were elicited through a single compound prompt, the observed differences are interpreted as discrepancies among observable outputs rather than evidence of independent cognitive stages or causal mechanisms.

The Recognition–Rendering Gap in Literary Metaphor Translation

The first major finding concerns the substantial discrepancy between GPT-5's recognition and interpretation scores and its Arabic rendering performance (RQ1). Recognition and interpretation achieved similar mean scores ($M = 2.30$ and 2.29 , respectively), whereas rendering received a considerably lower mean score ($M = 1.37$). This difference was statistically significant, with rendering scores lower than both recognition and interpretation scores. Furthermore, restricting the analysis to expressions receiving adequate or excellent recognition scores resulted in only a modest improvement in rendering performance.

These findings suggest that explicitly identifying an expression as metaphorical and explaining its contextual meaning do not necessarily coincide with an effective figurative realization of that expression in Arabic. This distinction is illustrated by the translation of *a welcome antidote to the noon sun*. Although GPT-5 identified the metaphor and explained its contextual meaning, its Arabic rendering communicated the general idea of relief without reproducing the medicinal imagery conveyed by *antidote*.

The finding is consistent with previous research showing that AI-generated literary translations may preserve propositional meaning while weakening figurative, cultural, or rhetorical features (Karakanta et al., 2025; Sahari et al., 2024; Shehab & Badah, 2026). However, the present study extends these observations by evaluating explicit metaphor identification and interpretation alongside final Arabic rendering. Rather than evaluating translation quality exclusively through the completed target text, the stage-aware framework makes it possible to identify cases in which relatively successful analytical outputs coexist with inadequate figurative rendering.

This distinction has implications for how metaphor translation is conceptualized and evaluated. Within cognitive approaches to metaphor, figurative expressions involve conceptual relationships that may be realized differently across languages (Kövecses, 2010; Lakoff & Johnson, 1980; Mandelblit, 1995). Successful literary translation therefore involves more than identifying a conceptual relationship or explaining its intended meaning. It also requires a target-language expression capable of preserving or reconstructing its relevant imagery, emotional resonance, and rhetorical function.

The results are compatible with Mandelblit's (1995) account of difficulties arising when metaphorical mappings differ across languages. Nevertheless, the present study did not directly measure mapping incongruence or the effort associated with translation. Accordingly, the observed recognition–rendering discrepancy should not be attributed exclusively to differences between English and Arabic conceptual mappings. Other factors, including lexical availability, textual context, rhetorical conventions, and prompting conditions, may also be relevant.

An important methodological consideration is that the three dimensions were elicited through the same standardized prompt. GPT-5 was explicitly instructed to identify metaphors, explain their meanings, and preserve figurative effects in Arabic. As a result, the recognition and interpretation outputs cannot be regarded as experimentally independent of the translation task. The

findings therefore do not establish whether the model recognized or understood a metaphor internally before generating its translation, nor do they show that a specific preceding difficulty caused the rendering failure.

Instead, the principal contribution lies in showing that explicit analytical performance is not a sufficient indicator of final literary translation quality under the prompting conditions investigated.

This distinction also highlights the limitations of relying exclusively on fluency or broad semantic equivalence when evaluating AI-generated literary translation. A grammatically acceptable and semantically intelligible Arabic expression may nevertheless alter or neutralize a metaphor's conceptual image. Such losses may be overlooked when evaluation prioritizes overall readability or propositional meaning rather than figurative realization.

The stage-aware framework therefore provides an additional evaluative perspective by making discrepancies among identification, explanation, and target-language expression visible. Its contribution is primarily diagnostic and methodological rather than an account of the internal processes through which GPT-5 generates translations.

Comparative Performance of GPT-5, DeepL, and Professional Translation

The second major finding concerns differences in rendering quality among GPT-5, DeepL, and the published Arabic translation (RQ2). DeepL achieved a moderately higher mean rendering score than GPT-5 ($M = 1.68$ vs. 1.37), while the published translation achieved a substantially higher mean score ($M = 2.55$). The differences among the three sources were statistically significant, and the published translation received the highest proportion of adequate or excellent renderings.

These findings are broadly consistent with previous research indicating that professional literary translations may preserve figurative imagery, cultural nuance, and stylistic effects more successfully than AI-generated alternatives in particular textual and linguistic contexts (Al Rousan et al., 2025; Manapbayeva et al., 2024; Zhang et al., 2025). They also resemble the findings of Karakanta et al. (2025), who reported a modest advantage for DeepL over GPT-4 in English–Dutch literary metaphor translation. However, these similarities should not be interpreted as evidence of a stable ranking across AI systems, literary genres, or language pairs.

The relatively small difference between GPT-5 and DeepL is noteworthy. Although the ordinal-score comparison favored DeepL, the proportion of adequate or excellent renderings differed by only 6.4 percentage points. This suggests that the two systems exhibited partially overlapping performance profiles, even though their distributions of rendering quality and failure types differed.

The qualitative findings provide additional insight into these patterns. DeepL demonstrated greater lexical consistency in some recurring expressions, as illustrated by its repeated use of الغيلان for *ogres*. GPT-5 employed several alternatives, including العفاريت, الكائنات المتوحشة, and الأوغار. Such variation may weaken textual continuity when recurring lexical items contribute to stable referential or imagistic patterns in the narrative.

Nevertheless, lexical consistency was not invariably associated with contextual accuracy. DeepL's rendering of *the boatman said* as قال القارب illustrates a referential error in which the person is confused with the vessel. This example demonstrates that lexical selection must be evaluated in relation to contextual meaning rather than through consistency alone.

The published professional translation displayed a different pattern, particularly in cases requiring rhetorical reformulation. The translation of *grey mane* as شعرها الأشيب مثل لبدة أسد illustrates how a metaphor may be reconstructed through an Arabic simile while preserving its central visual association. Both AI systems retained the descriptive reference to grey hair but did not reproduce the same animal-based image.

This example supports the argument that effective figurative equivalence does not necessarily depend on maintaining the source metaphor's original linguistic form. Rather, successful literary translation may involve selecting an alternative rhetorical construction that preserves the source expression's relevant conceptual and aesthetic effects (Abdul-Raof, 2006; Elewa, 2021; Schäffner, 2004).

The published translation's stronger performance should nevertheless be interpreted within the study's comparative design. The three translation sources were not produced under experimentally identical conditions. GPT-5 received explicit instructions concerning metaphor recognition, interpretation, and preservation, while DeepL was used as a dedicated translation system and the published translation represented an independently produced professional literary work.

Consequently, the observed differences establish relative quality within the evaluated translation products rather than the inherent superiority of human translators over AI systems.

The published translation was also not uniformly successful. The rendering of *She's got herself a candle* as حصلت لنفسها على شمعة illustrates a case in which literal lexical reproduction may provide limited access to broader symbolic meaning. However, whether such an instance constitutes metaphor loss depends on the linguistic classification of the source expression and should not be inferred from symbolic significance alone.

The findings therefore support a qualified conclusion: within the evaluated corpus, the published professional translation achieved stronger figurative rendering scores than the two AI systems. This advantage was particularly evident in examples requiring contextual integration and rhetorical reconstruction, but the extent to which it generalizes to other texts or translation conditions remains undetermined.

5.1 Linguistic and Translational Patterns Associated with Rendering Quality

The third major finding concerns the recurrent linguistic and translational patterns observed in successful and unsuccessful metaphor rendering (RQ3). Four patterns emerged from the qualitative analysis: lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation.

These patterns offer an interpretive account of variation in rendering quality, although they were not experimentally manipulated or independently measured as causal predictors.

Lexical equivalence and figurative meaning

Lexical equivalence was relevant when the source expression had an established or contextually appropriate Arabic counterpart. In such cases, relatively direct translation could preserve substantial elements of the source metaphor without extensive restructuring.

However, the findings also indicate that lexical correspondence and metaphorical equivalence are not identical. An Arabic translation may reproduce a source expression's descriptive meaning while losing the conceptual image that communicates it.

The *grey mane* example illustrates this distinction. Translating the expression as شعرها الرمادي preserves its reference to grey hair but neutralizes the comparison with an animal's mane. By contrast, the published translator's simile maintains the relevant visual relationship through a different rhetorical form.

This observation aligns with cognitive and functional approaches that emphasize conceptual correspondence and rhetorical effect over lexical identity alone (Lakoff & Johnson, 1980; Schäffner, 2004).

Discourse-Level Consistency and Narrative Continuity

The findings also highlight the relevance of discourse-level consistency in evaluating recurring expressions. Literary metaphors may function not only as isolated figurative constructions but also as elements contributing to narrative coherence, characterization, and the development of imagery across a text.

The original dataset showed considerable variation in the translation of the recurrent expression *boatman*. GPT-5 and DeepL employed different lexical and referential realizations, whereas the published translation consistently used ملاح.

Although the metaphorical classification of individual *boatman* occurrences requires verification, the example illustrates a broader translation-quality issue: inconsistency in rendering a recurring expression may affect textual continuity.

The analysis of *warren* also revealed a discrepancy between high reported recognition and interpretation scores and low rendering scores. GPT-5 frequently replaced the source expression with alternatives such as الممرات and المسكن, which conveyed aspects of the setting but did not preserve the original animal-burrow imagery.

Where *warren* is used metaphorically to conceptualize a network of human dwellings, these substitutions may weaken the underlying conceptual comparison. However, the interpretation of these instances depends on verifying the contextual and basic meanings required by MIP/MIPVU.

These examples suggest that metaphor-sensitive translation evaluation may benefit from examining discourse-level patterns rather than restricting analysis to isolated sentences.

Contextual Disambiguation and Conceptual Relationships

Contextual disambiguation emerged as another recurring pattern associated with translation quality. Certain expressions require translators to distinguish among possible lexical meanings, referents, or conceptual relationships using the surrounding narrative context.

The rendering of *the boatman said* as قال القارب illustrates an error of referential interpretation. Although the example is not necessarily a linguistic metaphor, it demonstrates how contextual misunderstanding can compromise translation quality.

More directly related to figurative meaning, the rendering of *my fading days* as an expression equivalent to *my remaining days* retains the broad idea of limited time while weakening the conceptual image of gradual fading.

Similarly, translating *our love withers* as يموت حبا preserves the general proposition that love is declining but modifies the source image. *Withering* conceptualizes deterioration through the gradual loss of vitality associated with plants, whereas *dying* emphasizes a different endpoint-oriented process.

Such substitutions are not inherently unacceptable. Their success depends on whether the resulting Arabic expression preserves the contextual implications and literary function of the original metaphor.

These findings reinforce the distinction between propositional accuracy and figurative equivalence. Semantic similarity may be sufficient to communicate the immediate situation without fully reproducing the metaphor's conceptual organization or rhetorical effect.

Rhetorical Recreation and Literary Effect

Rhetorical recreation was particularly prominent in examples where a relatively direct translation would not adequately reproduce the source expression's literary function.

In the *mask* sequence, the published translator employed صخرة صماء and later extended the image through تفلقت الصخرة وتدفت منها العواطف. These expressions establish a coherent progression from emotional rigidity to emotional release.

The published translation thus reconstructs the narrative's emotional development through a connected Arabic figurative pattern rather than relying solely on direct lexical correspondence. GPT-5 communicated the emotional content more explicitly but did not maintain the same progression of imagery.

This finding aligns with scholarship that emphasizes the importance of rhetorical adaptation in English–Arabic literary translation (Abdul-Raof, 2006; Elewa, 2021). It also supports the distinction between translating a metaphor's semantic content and reproducing its literary function.

Importantly, rhetorical recreation should not be equated automatically with higher translation quality. Creative reformulation may also introduce meanings or stylistic effects that depart from the source text. Its effectiveness must therefore be assessed against contextual accuracy, conceptual correspondence, and the requirements of the target literary text.

Taken together, the qualitative findings indicate that rendering quality depends on lexical, contextual, discourse-level, and rhetorical considerations. These factors help characterize the observed translation differences but do not establish a predictive model of translation success.

5.2 Theoretical and methodological implications

The findings have implications for both metaphor translation theory and evaluating AI-generated literary translations.

First, they reinforce the distinction between explicitly explaining metaphorical meaning and effectively realizing it in another language. In the present study, high recognition and interpretation scores were not consistently accompanied by high rendering scores. This suggests that evaluations focused exclusively on identifying or explaining figurative expressions may provide an incomplete account of literary translation quality.

Second, the findings support the need for metaphor-sensitive translation evaluation that considers contextual meaning, imagery, rhetorical effect, and discourse-level continuity. Existing approaches have emphasized the limitations of conventional evaluation measures in capturing figurative and literary qualities (Freitag et al., 2021; Karakanta et al., 2025; Wang et al., 2024). The present study complements this work by examining multiple observable outputs associated with a single translation task.

Third, the stage-aware framework provides a structured means of identifying discrepancies among metaphor detection, contextual explanation, and Arabic rendering. Its contribution lies in making these differences explicit and assessable rather than treating translation quality as a single undifferentiated outcome.

Nevertheless, the framework should not be interpreted as a validated model of sequential cognitive processing. The three outputs were generated in response to one compound prompt, and the study did not experimentally isolate the influence of recognition or interpretation on rendering quality. The observed associations therefore support diagnostic comparisons rather than causal explanations.

Fourth, the findings underscore the importance of defining metaphor consistently when constructing evaluation datasets. Linguistic metaphor, narrative symbolism, and allegorical reference may overlap in literary texts, but they are not equivalent analytical categories. The validity of metaphor-sensitive evaluation depends on applying identification criteria systematically and distinguishing linguistic metaphor from other literary phenomena.

Finally, the study suggests that evaluation frameworks for AI-generated literary translation could benefit from combining expression-level judgments with discourse-sensitive analysis. Recurrent imagery and referential continuity may influence the interpretation of individual renderings, particularly in narratives where figurative expressions contribute to broader thematic development.

Future validation of this approach could involve independently elicited prompting conditions, additional literary works, multiple language pairs, and statistical analyses accounting for repeated metaphor occurrences.

5.3 Limitations and directions for further research

Several limitations qualify the interpretation of the findings. The study examined 202 expressions from the first two chapters of a single novel, restricting its representativeness across literary genres, authors, and translation contexts. The findings may therefore reflect characteristics of Ishiguro's narrative style as well as the broader difficulties of English–Arabic literary metaphor translation.

In addition, metaphor identification relied on a predefined corpus. Recognition performance was assessed only for target expressions classified as metaphorical and did not include systematic measurement of false-positive identification. The distinction between linguistic metaphor and literary symbolism is particularly important in this context and warrants transparent documentation of borderline classifications.

GPT-5 was evaluated using one recorded response per passage through the ChatGPT web interface. We did not control model-generation parameters, and the design did not estimate variability across repeated outputs.

Recognition, interpretation, and rendering were also elicited using one compound prompt. Although this approach permitted comparison of three observable output dimensions, it did not establish their experimental independence or identify causal relationships among them.

The comparison with DeepL and the published professional translation involved different production conditions, limiting conclusions about the relative capabilities of translation systems or human translators beyond the specific outputs evaluated.

Furthermore, some expressions occurred repeatedly within the corpus. Consequently, individual observations may share lexical or contextual characteristics, potentially affecting the assumptions underlying statistical inference. Future research should examine whether the reported performance differences remain stable when repeated expressions are modeled or aggregated appropriately.

Finally, the rendering-quality framework involved holistic ordinal judgments. Although the study reported high inter-rater agreement, the operational boundaries among partial figurative preservation, substantial weakening, and complete de-metaphorization require continued refinement and validation across additional datasets.

Future research should therefore examine larger, more diverse literary corpora; independently elicited prompting conditions; repeated model generations; and evaluation designs that incorporate both metaphor-sensitive and discourse-level measures. Such research would help determine whether the observed recognition-rendering discrepancy holds across literary contexts and whether specific linguistic or rhetorical characteristics systematically predict rendering quality.

6. Conclusion

The findings indicate that GPT-5's strong performance in explicitly identifying and interpreting target metaphors did not consistently translate into effective Arabic figurative rendering. DeepL achieved moderately higher rendering scores, while the published professional translation performed substantially better than both AI systems within the evaluated corpus.

The qualitative analysis further suggests that lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation are important considerations in evaluating literary metaphor translation.

The study's principal contribution is developing and applying a stage-aware, output-level evaluation framework that reveals discrepancies between explicit metaphor explanation and target-language realization. Rather than identifying the internal causes of translation failure, this approach shows why evaluating literary metaphor requires attention to both conceptual meaning and its rhetorical reconstruction in the target language.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability: The datasets generated during the current study are available from the corresponding author upon reasonable request. The secondary literary source used in this study is publicly available through the published edition of *The Buried Giant* by Kazuo Ishiguro (2015) and its published Arabic translation by Kholoud Amr (2018).

Ethics approval: Not applicable.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Conflicts of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

- Abdul-Raof, H. (2006). *Arabic rhetoric: A pragmatic analysis*. Routledge.
- Al Rousan, R., Jaradat, R., & Malkawi, M. (2025). ChatGPT translation vs. human translation: An examination of a literary text. *Cogent Social Sciences*, 11(1), Article 2472916. <https://doi.org/10.1080/23311886.2025.2472916>
- Borowska-Szerszun, S. (2016). The giants beneath: Cultural memory and literature in Kazuo Ishiguro's *The Buried Giant*. *Crossroads: A Journal of English Studies*, 15, 30–41. <https://doi.org/10.15290/cr.2016.15.4.03>
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. <https://doi.org/10.1037/h0026256>

- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
- DeepL SE. (2025). DeepL Translator [Computer software]. <https://www.deepl.com/translator>
- Elewa, A. (2021). *Literary and cultural translation* (6th ed.). Qalam for Translation and Publication.
- Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9, 1460–1474. https://doi.org/10.1162/tacl_a_00437
- Gao, R., Lin, Y., Zhao, N., & Cai, Z. G. (2024). Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, 11, Article 835. <https://doi.org/10.1057/s41599-024-03363-0>
- Guerberof-Arenas, A., & Toral, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2), 184–212. <https://doi.org/10.1075/ts.21025.gue>
- Ishiguro, K. (2015). *The Buried Giant*. Faber & Faber.
- Ishiguro, K. (2018). *Al-‘imlāq al-madfūn [The Buried Giant]* (K. Amr, Trans.). Hamad Bin Khalifa University Press. (Original work published 2015)
- Karakanta, A., Nas, M., & Dorst, A. G. (2025). Metaphors in literary machine translation: Close but no cigar? In *Proceedings of Machine Translation Summit XX: Volume 1* (pp. 276–286). European Association for Machine Translation. <https://aclanthology.org/2025.mtsummit-1.21/>
- Kövecses, Z. (2010). *Metaphor: A practical introduction* (2nd ed.). Oxford University Press.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
- Levý, J. (2011). *The art of translation* (P. Corness, Trans.; Z. Jettmarová, Ed.). John Benjamins. <https://doi.org/10.1075/btl.97>
- Li, X., & Li, L. (2021). Reframed narrativity in literary translation: An investigation of the explicitation of cohesive chains. *Journal of Literary Semantics*, 50(2), 151–171. <https://doi.org/10.1515/jls-2021-2035>
- Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumática*, 12, 455–463. <https://doi.org/10.5565/rev/tradumatica.77>
- Manapbayeva, Z., Zaurbekova, G., Ayazbekova, K., Kazezova, A., & Pirmanova, K. (2024). AI in literary translation: ChatGPT-4 vs. professional human translation of Abai’s poem “Spring.” *Procedia Computer Science*, 251, 526–531. <https://doi.org/10.1016/j.procs.2024.11.143>
- Mandelblit, N. (1995). The cognitive view of metaphor and its implications for translation theory. In M. Thelen & B. Lewandowska-Tomaszczyk (Eds.), *Translation and meaning, Part 3* (pp. 483–495). Universitaire Press.
- Newmark, P. (1988). *A textbook of translation*. Prentice-Hall International.
- OpenAI. (2025). GPT-5 system card. <https://openai.com/index/gpt-5-system-card/>
- PACTE. (2003). Building a translation competence model. In F. Alves (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 43–66). John Benjamins.
- Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1), 1–39. <https://doi.org/10.1080/10926480709336752>
- Sahari, Y., Qasem, F., Asiri, E., Alasmri, I., Assiri, A., & Mahdi, H. (2024). Evaluating the translation of figurative language: A comparative study of ChatGPT and human translators. *CALR Linguistics Journal*, (15). <https://doi.org/10.60149/rtzq6644>
- Samaniego Fernández, E. (2022). Metaphor. In *ENTI: Encyclopedia of Translation and Interpreting*. AIETI. <https://doi.org/10.5281/zenodo.6369787>
- Schäffner, C. (2004). Metaphor and translation: Some implications of a cognitive approach. *Journal of Pragmatics*, 36(7), 1253–1269. <https://doi.org/10.1016/j.pragma.2003.10.012>
- Sharar, H., Al-Zagha, A. W., & Ali, E. (2025). ChatGPT in translating cultural nuances in “O Night” by Gibran Khalil Gibran. *International Journal of Linguistics, Literature and Translation*, 8(3), 30–38. <https://doi.org/10.32996/ijllt.2025.6.4.6>
- Shehab, E., & Badah, A. M. (2026). Translating conversational implicatures: A comparative study of human and artificial intelligence performance in Arabic–English literary translation. *Translation Review*, 124(1), 38–54. <https://doi.org/10.1080/07374836.2026.2617670>
- Steen, G. J. (1999). Analyzing metaphor in literature: With examples from William Wordsworth’s “I Wandered Lonely as a Cloud.” *Poetics Today*, 20(3), 499–522.
- Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). A method for linguistic metaphor identification: From MIP to MIPVU. John Benjamins. <https://doi.org/10.1075/celcr.14>
- Van den Broeck, R. (1981). The limits of translatability exemplified by metaphor translation. *Poetics Today*, 2(4), 73–87. <https://doi.org/10.2307/1772487>
- Wang, S., Zhang, G., Wu, H., Loakman, T., Huang, W., & Lin, C. (2024). MMTE: Corpus and metrics for evaluating machine translation quality of metaphorical language. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 11343–11358). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.634>
- Zhang, R., Zhao, W., & Eger, S. (2025). How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers* (pp. 10961–10988). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.548>