Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Hani Hamad M Albelihi

h.albelihi@qu.edu.sa

Qassim University

Arwa Alsaleh

Qassim University

Research Article

Keywords: literary metaphor translation, generative artificial intelligence, GPT-5, translation quality assessment, figurative language, English–Arabic translation

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11109228/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study introduces a stage-aware framework for evaluating GPT-5's Arabic translation of literary metaphor by separately assessing metaphor recognition, interpretation, and rendering. Using 202 metaphorical expressions identified through MIP/MIPVU in the first two chapters of Kazuo Ishiguro's *The Buried Giant*, the study compared GPT-5's recognition and interpretation outputs with its Arabic renderings and evaluated rendering quality against DeepL and Kholoud Amro's published translation. Each dimension was scored on a 0–3 ordinal scale, with high inter-rater reliability (weighted Cohen's κ = .89–.93). GPT-5 demonstrated strong metaphor recognition and interpretation (M = 2.30 and 2.29), but substantially weaker rendering performance (M = 1.37), with 37.6% of renderings scoring zero. DeepL achieved moderately higher rendering scores, while the published translation significantly outperformed both AI systems. Qualitative analysis identified four factors influencing rendering success: lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation. The findings reveal a substantial recognition–rendering gap, suggesting that conventional fluency- or comprehension-based evaluations may overestimate AI performance in literary metaphor translation. The study offers a replicable methodology for diagnosing AI limitations in figurative language processing.

Introduction

The quality of metaphor translation is regarded as a key aspect of evaluating literary translation (Levý, 2011). The construct enables evaluators to assess how well a translation preserves the conceptual, emotional, and rhetorical functions of the source metaphor, providing a more detailed measure of translation adequacy than fluency alone (Newmark, 1988; van den Broeck, 1981; Schäffner, 2004). Despite their growing use in literary translation contexts, AI-assisted translation systems still struggle to render complex metaphors (Guerberof-Arenas & Toral, 2022; Manapbayeva et al., 2024; Zhang et al., 2025). Large language models (LLMs) such as GPT-5 (OpenAI, 2025) can produce fluent translations, but existing evaluation approaches may not adequately reveal whether figurative meaning, imagery, and conceptual relationships have been preserved in the target text (Freitag et al., 2021; Wang et al., 2024; Karakanta et al., 2025).

To address this problem, the present study focuses on recognition, interpretation, and rendering as analytically distinct dimensions of literary metaphor translation. Recognition refers to identifying an expression as metaphorical (PACTE, 2003), while interpretation concerns construing its intended conceptual and emotional meaning (Mandelblit, 1995), including the relationships between source and target domains and the affective implications of the figurative expression (Lakoff & Johnson, 1980; Kövecses, 2010; Mandelblit, 1995; Schäffner, 2004; Abdul-Raof, 2006). Translation studies have widely evaluated AI-generated literary translation through the completed target text, with previous research showing that fluency-oriented or general translation-quality measures may overlook losses in figurative meaning (Freitag et al., 2021; Wang et al., 2024; Karakanta et al., 2025). Nonetheless, few empirical

studies have separately evaluated metaphor recognition and interpretation alongside the final rendering (Mandelblit, 1995; Karakanta et al., 2025).

This limitation is important because existing metaphor-sensitive frameworks principally evaluate completed translations (Wang et al., 2024; Karakanta et al., 2025). From an unsuccessful target-language rendering alone, it is difficult to determine whether the model failed to identify the metaphor, produced an inadequate interpretation of its contextual meaning, or struggled to express that meaning figuratively in the target language. The present study addresses this limitation by eliciting recognition, interpretation, and rendering separately, allowing each dimension of GPT-5's observable output to be evaluated independently. The completed rendering then provides a common basis for comparison with DeepL and a published professional translation, for which recognition and interpretation cannot be directly elicited.

For instance, Karakanta et al. (2025) reported divergence between automatic and human judgments of literary metaphor translation, underscoring the limitations of evaluating figurative language solely through finished-text measures. Building on this concern, we introduce a stage-aware framework for evaluating GPT-5's Arabic translations of literary metaphor. Here, stage-aware refers to separately evaluating elicited recognition, interpretation, and rendering outputs rather than making claims about the model's internal cognitive processing. This analytical approach is intended to identify whether weaknesses in the final rendering of metaphors are associated with recognition, interpretation, or target-language realization, and to examine the linguistic and translational factors that characterize successful and unsuccessful outputs.

Much recent research on AI-assisted literary translation and metaphor translation has examined language pairs other than English–Arabic (Karakanta et al., 2025; Gao et al., 2024; Manapbayeva et al., 2024; Zhang et al., 2025). Consequently, relatively little is known about AI metaphor translation in a pair characterized by substantial differences in linguistic structure, rhetorical conventions, and figurative traditions (Abdul-Raof, 2006; Samaniego Fernández, 2022; Schäffner, 2004). English–Arabic therefore provides a useful context for examining how literary imagery and figurative meaning are maintained, modified, or lost across translation. We designed a stage-aware comparative analysis of 202 literary metaphors from Kazuo Ishiguro's *The Buried Giant* (Ishiguro, 2015), using separately elicited GPT-5 recognition, interpretation, and rendering outputs alongside completed Arabic translations produced by DeepL and Kholoud Amro's published professional translation.

Literature Review

Metaphor translation

Earlier approaches to metaphor translation centered on the procedures available to translators, preserving the source image, replacing it with another metaphor, resorting to literal paraphrase, or omitting the figure altogether (Newmark, 1988; van den Broeck, 1981; Schäffner, 2004). Cognitive

approaches later reframed metaphor as a conceptual mapping that varies across languages and cultures, shifting the analytic focus from procedure to process (Lakoff & Johnson, 1980; Kövecses, 2010; Schäffner, 2004). This process-oriented turn provides a basis for treating recognition, interpretation, and rendering as separate analytical dimensions in the present study (PACTE, 2003; Mandelblit, 1995). 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).

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 (Schäffner, 2004; Samaniego Fernández, 2022; Abdul-Raof, 2006). 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; Schäffner, 2004; Samaniego Fernández, 2022). Such changes are particularly consequential 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.

Toward metaphor-sensitive evaluation

Traditional machine-translation metrics are of limited usefulness for evaluating literary metaphors because they primarily focus 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; Wang et al., 2024; Karakanta et al., 2025). 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, on the

grounds that mainstream metrics give 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 (Wang et al., 2024; Freitag et al., 2021; Karakanta et al., 2025). 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 (Mandelblit, 1995; Karakanta et al., 2025).

Variation across LLM, NMT, and human translation

No consensus yet exists on the superiority of any single AI system in 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. Gao et al. (2024), for instance, revealed that ChatGPT's translations of Chinese classical poetry improved once 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 particular 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 (Zhang et al., 2025; Manapbayeva et al., 2024; Al Rousan et al., 2025; Sahari et al., 2024). Within this pair, Al Rousan, Jaradat, and Malkawi 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 relationship between separately elicited recognition, interpretation, and rendering outputs, and the behavior of a specialized NMT system are 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 successful interpretation of figurative meaning and successful target-language rendering 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 (Freitag et al., 2021; Wang et al., 2024; Karakanta et al., 2025; Mandelblit, 1995; Al Rousan et al., 2025).

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 Amro's published translation (Ishiguro, 2018). Recognition and interpretation are assessed only for GPT-5 because they cannot be directly elicited from DeepL or reliably reconstructed 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 Amro'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?

Methodology

This study used a mixed-methods design to investigate the recognition, interpretation, and Arabic rendering of literary metaphors. Quantitative analysis was used to examine differences across GPT-5's

three analytical outputs and to compare the metaphor renderings produced by GPT-5, DeepL, and a published professional translation. Qualitative textual analysis was used to further examine recurrent patterns associated with successful and unsuccessful metaphor rendering (Creswell & Plano Clark, 2018).

Corpus and translation sources

The Buried Giant (Ishiguro, 2015) was chosen for the corpus for textual and translational reasons. Its narrative is rich in metaphor, symbolism, allegory, and imagery, making it well suited for metaphor analysis (Borowska-Szerszun, 2016; Li & Li, 2021; Karakanta et al., 2025). The first two chapters allowed examination of metaphors in context and tracking recurring imagery, covering 202 expressions. Additionally, the novel's Arabic translation (Ishiguro, 2018) by Kholoud Amrou, who won the Third Prize in the Sheik Hamad Award for Translation, provides a professional standard for comparing AI-generated Arabic metaphors.

Three translation sources were used. GPT-5 produced the metaphor identifications, contextual interpretations, and Arabic translations. DeepL was used for the English-to-Arabic translations, and Kholoud Amro's published Arabic version of The Buried Giant was included as a comparative example from a professional literary source. Although the published translation was not considered completely error-free, it was treated as an established human translation and thus served as a basis for comparing the AI-generated translations.

Metaphor identification

The source text was coded for metaphorical expressions using the Metaphor Identification Procedure (MIP) developed by the Pragglejaz Group (Pragglejaz Group, 2007), supplemented by principles from MIPVU (Steen et al., 2010). Each lexical unit was examined in context by comparing its contextual meaning with a more basic meaning. We identified a lexical unit as metaphorically used when the two meanings contrasted but could be understood in relation to one another.

Although metaphor identification was conducted at the lexical-unit level, the complete metaphorical expression was retained for subsequent translation analysis. This approach enabled consideration of structural differences between English and Arabic, ensuring that word-level correspondence did not solely determine translation quality. We independently applied the identification procedure. Agreement was 92%, and disagreements were resolved through discussion before the final corpus of 202 metaphors was established. This corpus was then used for subsequent data collection and analysis.

Data collection

GPT-5 was accessed through the OpenAI ChatGPT web interface on 13 October 2025 (OpenAI, 2025). The interface did not expose model-generation parameters, such as temperature, so they were neither

independently controlled nor recorded. For each target passage, the complete source text (ST) was submitted in a fresh conversation, allowing GPT-5 to interpret the target metaphors within the broader narrative context while minimizing contextual carryover from previously analyzed items. A single response was collected for each passage, and no regeneration or manual selection among alternative outputs was performed. The findings therefore represent the set of model outputs generated under these conditions rather than an estimate of variability across repeated generations.

A standardized prompt was used to generate 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.

GPT-5's responses remained unchanged. 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 since similar outputs from DeepL were unavailable and could not be reliably recreated from the published translation. In contrast, rendering was evaluated for all three sources.

Each metaphor was recorded in an Excel dataset along with its source sentence, relevant narrative context, GPT-5 recognition and interpretation results, GPT-5 and DeepL translations, and the related segment from Amro's translation. The alignment was performed on the full metaphorical expression rather than through word-by-word matching.

Evaluation and coding

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

The rendering score represented an overall judgment rather than separate numerical scores for each component. Semantic accuracy and preservation of metaphorical meaning were treated as necessary conditions for scores of 2 or 3, while naturalness, imagery, tone, style, and figurative force were used to distinguish adequate from excellent renderings. A rendering that preserved the general propositional meaning but substantially weakened or neutralized the metaphor was normally assigned a score of 1. Renderings that distorted the contextual meaning or removed the metaphor entirely were assigned a score of 0 according to the 0T and 0M categories described below. Reformulations were not penalized if they maintained the metaphor’s meaning and literary impact through another suitable rhetorical form.

The three categories were treated as analytical dimensions of observable model output rather than as evidence of sequential internal processing. Because recognition was assessed using preidentified metaphorical expressions, the recognition scores reflected GPT-5's detection of target metaphors rather than unrestricted metaphor-classification performance. The design therefore did not assess false-positive identification of non-metaphorical expressions.

Two types of zero-score rendering were distinguished: mistranslation (0T), in which the contextual meaning was distorted, and omission/de-metaphorization (0M), in which the figurative image was omitted or neutralized. A score of 1 was used when part of the source metaphor remained despite substantial weakening.

We independently coded 202 recognition outputs, 202 interpretation outputs, and 606 rendering observations, comprising 202 renderings from each of GPT-5, DeepL, and the published translation. Weighted Cohen's κ was calculated separately for recognition, interpretation, and rendering because the ratings were ordinal (Cohen, 1968). For rendering, the reliability analysis pooled the 606 observations across the three translation sources. Percentage agreement was 91% for recognition, 88% for interpretation, and 89% for rendering, with weighted κ values of .93, .89, and .91, respectively. Disagreements were resolved through discussion.

Table 1
Operational scoring framework for metaphors

Score	Recognition	Interpretation	Rendering
3: Excellent	The target expression is clearly identified as metaphorical	Explanation accurately captures the metaphor's meaning, relationship, and function.	Metaphor translates well into Arabic, preserving its image, meaning, tone, and figurative power.
2: Adequate	The target expression is correctly identified as metaphorical, though the identification may be less explicit	The explanation is generally accurate but has a minor omission or a reduction in nuance.	Rendering is accurate and natural but loses some imagery, tone, style, or figurative force.
1: Weak	The response shows partial or uncertain recognition of the target metaphor.	Explanation is vague, incomplete, overly literal, or poorly connected to the narrative.	Rendering has weaknesses, possibly too literal, awkward, or flat, but the overall meaning stays clear.
0: Failure	The target metaphor is not identified as metaphorical.	No relevant interpretation is provided, or explanation distorts the metaphor's meaning.	rendering either mistranslates the metaphor (0T) or omits, neutralizes, or de-metaphorizes it (0M).

As shown in Table 2, we employed complementary methods of data analysis: (1) descriptive statistical analysis, (2) nonparametric inferential analysis, and (3) qualitative textual analysis.

To answer RQ1, descriptive statistics were first used to characterize GPT-5's performance in recognition, interpretation, and rendering. Frequencies, percentages, means, medians, standard deviations, and rates of adequate or excellent performance were calculated. Because the same 202 metaphorical expressions received ordinal scores for recognition, interpretation, and rendering, a Friedman test was used to examine differences across the three dimensions. Significant overall differences were followed by Holm-adjusted Wilcoxon signed-rank tests. Kendall's *W* was reported as the effect size for the Friedman test, while $r = |Z|/\sqrt{N}$ was used for pairwise comparisons. For pairwise comparisons, Holm-adjusted *p-values* and effect sizes were reported.

To answer RQ2, the same procedures were applied to the rendering scores of GPT-5, DeepL, and the published Arabic translation. Frequencies and percentages of mistranslation (0T) and omission/de-metaphorization (0M) were also calculated to characterize differences in failure patterns across the three translation sources.

To answer RQ3, qualitative analysis was conducted alongside quantitative coding to identify recurring linguistic and translational patterns associated with successful and unsuccessful metaphor rendering. The analysis was guided by established metaphor-translation concepts such as de-metaphorization and conceptual substitution, with additional patterns identified through repeated examination of the data. These included lexical transfer, contextual disambiguation, discourse-level consistency, and rhetorical recreation. Cases with high and low rendering scores were examined within their narrative context and compared across GPT-5, DeepL, and the published translation. Examples were selected to illustrate recurring patterns rather than isolated errors.

Table 2
Methods of data analysis

Data source	**RQ1**	**RQ2**	**RQ3**
	GPT-5 recognition, interpretation, and rendering	**GPT-5, DeepL, and published renderings**	**Recurrent rendering cases**
Quantitative analysis	Descriptive statistics; Friedman; Wilcoxon; effect sizes	Descriptive statistics; Friedman; Wilcoxon; failure frequencies	—
Qualitative analysis	—	—	Textual analysis of recurrent patterns

Results

This section has three parts aligned with the research questions. RQ1 reports GPT-5's performance in recognition, interpretation, and rendering. RQ2 compares the metaphor rendering of GPT-5, DeepL, and the published Arabic translation. RQ3 examines linguistic and translational factors linked to rendering performance.

GPT-5’s performance across recognition, interpretation, and rendering

Table 3 summarizes GPT-5’s performance across the three analytical dimensions. Recognition and interpretation produced nearly identical mean scores (*M* = 2.30 and 2.29), with 77.2% and 76.7% of cases rated adequate or excellent (≥ 2), respectively. Rendering was substantially weaker (*M* = 1.37, *Mdn* = 1, *SD* = 1.22), with only 50.0% of renderings rated adequate or excellent. Zero scores occurred in 37.6% of renderings, including 22.8% omission or de-metaphorization (0M) and 14.9% mistranslation (0T).

Table 3
GPT-5’s performance across metaphor recognition, interpretation, and rendering

Dimension	***n***	***M***	***Mdn***	***SD***	***≥ 2 (%)***	***0 (%)***	***0M (%)***	***0T (%)***
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

Recognition and interpretation were closely aligned, whereas higher performance on these dimensions did not consistently correspond to successful Arabic rendering. The expression *a welcome antidote to the noon sun* illustrates this pattern. GPT-5 identified *antidote* as metaphorical and interpreted the wind as relief from oppressive heat, but rendered it as:

The propositional meaning was retained, but the medicinal image was lost.

To examine whether weak rendering was mainly associated with unsuccessful recognition, rendering scores were recalculated for the 156 metaphors with recognition scores of 2 or 3. The mean rendering score increased only from 1.37 to 1.51, and 56.4% of these cases were rated adequate or excellent. Thus, low rendering performance was not confined to cases in which the preidentified target metaphor was not successfully recognized.

Inferential analysis confirmed the differences across the three dimensions (Table 4). The Friedman test showed a significant overall difference among recognition, interpretation, and rendering, $\chi^2(2, N = 202) = 108.95$, $p < .001$, Kendall’s $W = .27$. Holm-adjusted Wilcoxon signed-rank tests found no significant difference between recognition and interpretation (adjusted $p = .32$), whereas rendering was significantly lower than both recognition and interpretation (both adjusted ps $< .001$, $r = .65$).

Table 4
Inferential comparisons among recognition, interpretation, and rendering (GPT-5)

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

These findings answer RQ1 by showing that the largest differences were observed between rendering and the recognition and interpretation dimensions, while recognition and interpretation remained closely aligned.

Comparative metaphor-rendering performance

Table 5 compares the rendering performance of GPT-5, DeepL, and the published Arabic translation. The published 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, 56.4% of DeepL, and 50.0% of GPT-5.

Table 5
Metaphor-rendering performance by output source

Output source	*n*	*M*	*Mdn*	*SD*	*≥ 2 (%)*	*0 (%)*	*0M (%)*	*0T (%)*
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

DeepL received significantly higher rendering scores than GPT-5 (adjusted *p* < .001, *r* = .34), although the difference in adequate-or-excellent rendering rates was relatively small. The qualitative examples indicate greater lexical consistency in some DeepL renderings. For example, DeepL consistently translated *ogres* as □□□□□□□, whereas GPT-5 varied among □□□□□□□□ ,□□□□□□□□ □□□□□□□□ and □□□□□□□. However, this consistency did not always correspond to successful contextual interpretation. DeepL rendered *the boatman said* as □□□□□□ □□□, confusing the person with the vessel.

The overall comparison among the three translation sources was significant, $\chi^2(2, N = 202) = 127.23$, *p* < .001, Kendall's *W* = .32 (Table 6). Holm-adjusted Wilcoxon signed-rank tests showed significant differences between GPT-5 and DeepL (adjusted *p* < .001, *r* = .34), between GPT-5 and the published translation (adjusted *p* < .001, *r* = .80), and between DeepL and the published translation (adjusted *p* < .001, *r* = .67). The highest score of 3 was assigned to 79.2% of the published renderings, compared with 34.7% of DeepL renderings and 24.3% of GPT-5 renderings.

Table 6
Inferential comparisons among the three translation sources

Comparison	Holm-adjusted *p*	Effect size
Omnibus Friedman: χ^2 (2) = 127.23	< .001	Kendall's *W* = .32
GPT-5 vs. DeepL	< .001	*r* = .34
GPT-5 vs. Published translation	< .001	*r* = .80
DeepL vs. Published translation	< .001	*r* = .67

The *grey mane* example illustrates a difference in rhetorical treatment. GPT-5 and DeepL rendered the expression as [redacted], preserving its descriptive meaning while removing the animal imagery. The published translator used [redacted], reformulating the metaphor as a simile while retaining its visual force.

Failure profiles also differed across the three sources. Zero scores accounted for 37.6% of GPT-5 renderings, 23.3% of DeepL renderings, and 10.9% of the published renderings. Mistranslation (0T) accounted for 14.9% of GPT-5 and 8.4% of DeepL renderings, whereas none of the published translations' zero-score cases were classified as mistranslations. Omission or de-metaphorization (0M) accounted for 22.8%, 14.9%, and 10.9% of GPT-5, DeepL, and published renderings, respectively.

The published translation was nevertheless not uniformly successful. *"She's got herself a candle" was translated literally as* [redacted]*, which only partially conveys* its symbolic significance. It was therefore treated as a professional comparator rather than an error-free gold standard. Overall, the findings for RQ2 show that DeepL achieved moderately higher rendering scores than GPT-5, while the published translation achieved substantially higher scores than both AI systems in this corpus.

Linguistic and translational factors affecting metaphor rendering

Closer textual analysis identified four recurrent factors associated with rendering performance: lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation. Table 7 illustrates these patterns across five recurring metaphors.

Table 7
Mean scores for five recurring metaphors by dimension and output source.

Metaphor	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

Note. Scores ranged from 0 to 3. Rec. = recognition; Int. = interpretation; Ren. = rendering. The result for the buried giant metaphor is based on only two occurrences and should therefore be interpreted cautiously.

Table 7 shows stronger rendering for *mist* and *candle* than for *warren* and *boatman*, indicating variation in line with the contextual and discourse-level demands of the metaphors.

The *warren* metaphor provides the clearest contrast between recognition, interpretation, and rendering. GPT-5 achieved mean scores of 3.00 for both recognition and interpretation across all ten occurrences, but a rendering mean of 0.00. It replaced the source image with expressions such as [illegible] and [illegible], conveying aspects of the setting while eliminating the image of an interconnected network of animal burrows. This pattern illustrates how successful recognition and interpretation did not necessarily correspond to preservation of the metaphorical image in Arabic.

Discourse-level consistency and contextual disambiguation also affected rendering. For *boatman*, the published translator consistently used [illegible], whereas GPT-5 and DeepL varied among [illegible] ,[illegible], pronouns, and, [illegible]. The latter reflected a contextual misinterpretation of the referent, while the broader variation weakened lexical continuity across the narrative. DeepL nevertheless achieved a higher rendering mean than GPT-5 for this metaphor (*M* = 1.62 vs. 0.69).

De-metaphorization and conceptual substitution were also recurrent. GPT-5 conveyed *that the birds were now in chorus*, as if singing in unison, removing the image of coordinated musical performance and rendering *my fading days* as *my remaining days*, preserving temporal meaning while weakening the image of gradual loss of vitality. Conceptual substitution produced a related shift when *our love withers* became [illegible] [illegible]. Although both expressions convey decline, *withering* presents a gradual process, whereas *dying* introduces a different endpoint-oriented image.

The largest differences appeared when successful rendering required rhetorical recreation rather than direct lexical transfer. In the *mask* sequence, the published translator used [illegible] [illegible] and later extended the image through [illegible] [illegible] [illegible] [illegible] [illegible], creating a coherent figurative progression from

emotional rigidity to release. GPT-5 conveyed the emotional meaning more explicitly but did not maintain the same figurative development.

Overall, the findings for RQ3 show that successful literary rendering was associated with four recurrent factors: appropriate lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation. Weaknesses in these areas were associated with de-metaphorization, conceptual substitution, inconsistency, and loss of figurative continuity.

Discussion

The gap between recognition and rendering in translating literary metaphors

The first major finding of this study is the substantial difference between GPT-5's performance in metaphor recognition and interpretation and its performance in Arabic rendering (RQ1). Recognition and interpretation remained closely aligned, whereas rendering was substantially weaker. Even when the analysis was restricted to metaphors that GPT-5 successfully recognized, rendering performance improved only slightly. These findings indicate that weak Arabic rendering was not confined to cases where the preidentified target metaphor was unrecognized.

This finding builds on previous work on AI-generated literary translation by more precisely locating metaphorical difficulty. Studies of LLMs and NMT have shown that fluent translations can still lose figurative and literary information (Karakanta et al., 2025; Sahari et al., 2024). The present findings extend this observation by showing that such loss can occur even when GPT-5 produces an adequate explicit interpretation of the source metaphor. In this respect, the results are also consistent with Shehab and Badah (2026), who found that AI-generated literary translations may preserve surface meaning while failing to reproduce pragmatic force as successfully as professional translation.

The distinction is important because evaluation based solely on the completed translation cannot readily indicate whether an unsuccessful rendering is due to inadequate recognition of metaphor, inaccurate interpretation, or difficulty with target-language realization. The stage-aware framework used here makes these differences observable across separately elicited model outputs without treating them as evidence of sequential internal processing. In the present data, the largest differences occurred between rendering and the recognition and interpretation dimensions.

Recurring metaphors reinforced this pattern: *warren* and *boatman* received high recognition and interpretation scores but substantially weaker renderings. This confirms that the explicit interpretation of metaphor does not necessarily correspond to successful figurative reconstruction in Arabic.

The recognition–rendering gap also has implications for evaluating the quality of LLM translations. A semantically intelligible and grammatically fluent translation may still weaken or remove the figurative structure through which literary meaning is produced. Evaluation based largely on fluency or

propositional equivalence may therefore overestimate performance in literary metaphor translation. Metaphor-sensitive evaluation needs to consider how figurative meaning is realized in the target text, not simply whether the general message remains recoverable.

AI systems and professional literary metaphor translation

The second major finding concerns differences among GPT-5, DeepL, and the published Arabic translation (RQ2). DeepL achieved moderately higher rendering scores than GPT-5, but the difference between the two AI systems was considerably smaller than the differences between either system and the published translation. This result is broadly consistent with Karakanta et al. (2025), who reported a modest advantage for DeepL over GPT-4 in English–Dutch literary metaphor translation, although comparative performance should not be generalized across systems, tasks, and language pairs. Gao et al. (2024) similarly showed that LLM literary-translation performance can vary when prompts explicitly foreground literary features.

The qualitative evidence helps explain some of the differences between the two AI systems. DeepL showed greater lexical consistency when established Arabic equivalents were available. Its consistent rendering of *ogres* as □□□□□□□ contrasts with GPT-5's use of several different alternatives. Such consistency can contribute to continuity when an expression recurs across a literary text. However, DeepL's mistranslation of *boatman* as □□□□ demonstrates that lexical stability alone cannot compensate for incorrect contextual disambiguation.

The published translation showed a different pattern. Its advantage in this corpus was associated not simply with grammatical fluency but with a stronger combination of contextual interpretation, lexical continuity, and rhetorical adaptation. This finding is consistent with Zhang et al. (2025); their comparison also identified advantages of professional literary translation over LLM output. The grey mane example illustrates the nature of this difference. GPT-5 and DeepL rendered the expression as □□□□□□□ □□□□□, retaining its descriptive meaning but removing the animal image. The published translator instead used □□□ □□□□ □□□ □□□□□□ □□□□□, reformulating the metaphor as a simile while preserving its visual effect. This example supports the view that figurative equivalence may require rhetorical reconstruction rather than direct lexical correspondence (Abdul-Raof, 2006; Elewa, 2021).

The published translation was nevertheless not uniformly successful. The literal rendering of *She's got herself a candle* weakened part of the symbolic meaning, showing that professional translation should not be treated as an error-free gold standard. The findings therefore do not establish general human superiority. Rather, they show that this published professional translation achieved substantially stronger metaphor-rendering performance than GPT-5 and DeepL within the present corpus.

Factors shaping successful metaphor rendering

The third major finding concerns the recurrent linguistic and translational factors associated with successful and unsuccessful rendering (RQ3). Four factors were particularly evident across the data: lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation.

Their influence varied according to the type of metaphor and the degree of contextual or rhetorical reconstruction required.

Relatively direct lexical transfer was more successful when a suitable and stable Arabic equivalent was available. However, lexical correspondence became less reliable when expressions depended on wider narrative context. The *boatman* examples demonstrate this clearly: several lexical alternatives were individually plausible, but variation weakened referential continuity, while the rendering □□□□ represented a contextual misinterpretation of the referent. This suggests that literary metaphor rendering cannot be evaluated solely at the sentence level, because repeated expressions may contribute to patterns that extend across the narrative.

De-metaphorization and conceptual substitution illustrate a related problem. Renderings such as *my fading days* as *my remaining days* retained the general propositional meaning but weakened the figurative image. Likewise, □□□□ □□□□ for *our love withers* retained the broader idea of decline while replacing the gradual image of withering with a different conceptual endpoint. Such cases show that semantic similarity does not necessarily entail metaphorical equivalence.

Some of the strongest published renderings involved rhetorical recreation rather than literal transfer. In the *mask* sequence, the published translator rendered □□□□ □□□□ as □□□□□□□ □□□□ □□□□□□ □□□□□□ □□□□□, maintaining a coherent figurative progression in a form that functions naturally in Arabic. This pattern is consistent with scholarship on English–Arabic literary translation that emphasizes rhetorical reconstruction rather than the mechanical reproduction of source-language form (Abdul-Raof, 2006; Elewa, 2021).

Taken together, the findings suggest that successful literary metaphor rendering depends on more than preserving propositional meaning. It requires attention to lexical choice, contextual reference, recurrence across discourse, conceptual imagery, and target-language rhetorical resources. The stage-aware framework contributes to this analysis by showing that weak rendering can coexist with strong recognition and interpretation, while the cross-source comparison demonstrates how different translation systems respond to these demands in distinct ways.

Conclusion

This mixed-methods study examined recognition, interpretation, and rendering in the English–Arabic translation of 202 literary metaphors from Kazuo Ishiguro's *The Buried Giant*. GPT-5 performed strongly in recognizing and interpreting the preidentified metaphors but substantially less successfully in rendering them into Arabic. DeepL achieved moderately higher rendering scores than GPT-5, while Kholoud Amro's published translation achieved substantially higher scores than both AI systems in this corpus. The analysis also identified four recurrent factors associated with rendering quality: lexical equivalence, discourse-level consistency, contextual disambiguation, and rhetorical recreation.

These findings demonstrate the value of stage-aware evaluation in literary translation. Separately eliciting recognition, interpretation, and rendering makes it possible to identify differences across observable model outputs without treating them as evidence of internal cognitive processing. The findings further suggest that preserving propositional meaning alone is insufficient for translating literary metaphor; successful rendering also requires attention to imagery, lexical continuity, context, and rhetorical effect.

Several limitations should be acknowledged. The study examined only the first two chapters of a single novel and a single translation direction. Recognition was assessed only for preidentified metaphors and therefore did not measure false-positive identification. GPT-5 outputs were generated once via the web interface without control over sampling parameters, and prior exposure to the novel or published Arabic translation cannot be ruled out. Future research should examine larger literary corpora, additional language pairs and model versions, and repeated generations to determine whether the recognition–rendering gap persists across contexts.

Declarations

Funding declaration:
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 Amro (2018).

Ethics declaration:
Not applicable.

Consent to Participate

Not Applicable

Consent for Publication

Not Applicable.

Declaration of Conflicting Interests:
The authors declare no potential conflicts of interest regarding the research, authorship, or publication of this article.

Acknowledgments:

The authors would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2026).

References

1. Abdul-Raof H. Arabic rhetoric: A pragmatic analysis. Routledge; 2006.
2. Al Rousan R, Jaradat R, Malkawi M. ChatGPT translation vs. human translation: An examination of a literary text. Cogent Social Sci. 2025;11(1). https://doi.org/10.1080/23311886.2025.2472916. Article 2472916.
3. 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
4. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20. https://doi.org/10.1037/h0026256.
5. Creswell JW, Plano Clark VL. Designing and conducting mixed methods research. 3rd ed. SAGE; 2018.
6. DeepL SE. (2025). *DeepL Translator* [Computer software]. https://www.deepl.com/translator
7. Elewa A. Literary and cultural translation. 6th ed. Qalam for Translation and Publication; 2021.
8. Freitag M, Foster G, Grangier D, Ratnakar V, Tan Q, Macherey W. Experts, errors, and context: A large-scale study of human evaluation for machine translation. Trans Association Comput Linguistics. 2021;9:1460–74. https://doi.org/10.1162/tacl_a_00437.
9. Gao R, Lin Y, Zhao N, Cai ZG. Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. Humanit Social Sci Commun. 2024;11:835. https://doi.org/10.1057/s41599-024-03363-0.
10. Guerberof-Arenas A, Toral A. Creativity in translation: Machine translation as a constraint for literary texts. Translation Spaces. 2022;11(2):184–212. https://doi.org/10.1075/ts.21025.gue.
11. Ishiguro K. The Buried Giant. Faber & Faber; 2015.
12. Ishiguro K. Trans). Al-ʾimlāq al-madfūn [The Buried Giant]. Hamad Bin Khalifa University; 2018. (Original work published 2015). K. Amro.
13. Karakanta A, Nas M, Dorst AG. (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/.
14. Kövecses Z. Metaphor: A practical introduction. 2nd ed. Oxford University Press; 2010.
15. Lakoff G, Johnson M. Metaphors we live by. University of Chicago Press; 1980.
16. Levý J. In: Jettmarová Z, editor. The art of translation. Trans.; 2011. https://doi.org/10.1075/btl.97. P. CornessJohn Benjamins.
17. Li X, Li L. Reframed narrativity in literary translation: An investigation of the explicitation of cohesive chains. J Literary Semant. 2021;50(2):151–71. https://doi.org/10.1515/jls-2021-2035.

18. Lommel A, Uszkoreit H, Burchardt A. Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica. 2014;12:455–63. https://doi.org/10.5565/rev/tradumatica.77.
19. Manapbayeva Z, Zaurbekova G, Ayazbekova K, Kazezova A, Pirmanova K. AI in literary translation: ChatGPT-4 vs. professional human translation of Abai's poem Spring. Procedia Comput Sci. 2024;251:526–31. https://doi.org/10.1016/j.procs.2024.11.143.
20. Mandelblit N. The cognitive view of metaphor and its implications for translation theory. In: Thelen M, Lewandowska-Tomaszczyk B, editors. Translation and meaning, Part 3. Universitaire; 1995. pp. 483–95.
21. Newmark P. A textbook of translation. Prentice-Hall International; 1988.
22. OpenAI. (2025). *GPT-5 system card*. https://openai.com/index/gpt-5-system-card/
23. PACTE. Building a translation competence model. In: Alves F, editor. Triangulating translation: Perspectives in process oriented research. John Benjamins; 2003. pp. 43–66.
24. Pragglejaz Group. MIP: A method for identifying metaphorically used words in discourse. Metaphor Symbol. 2007;22(1):1–39. https://doi.org/10.1080/10926480709336752.
25. Sahari Y, Qasem F, Asiri E, Alasmri I, Assiri A, Mahdi H. Evaluating the translation of figurative language: A comparative study of ChatGPT and human translators. Research Square [Preprint]. 2024. https://doi.org/10.21203/rs.3.rs-3921149/v1.
26. Samaniego Fernández E. (2022). Metaphor. In *ENTI: Encyclopedia of Translation and Interpreting*. AIETI. https://doi.org/10.5281/zenodo.6369787
27. Schäffner C. Metaphor and translation: Some implications of a cognitive approach. J Pragmat. 2004;36(7):1253–69. https://doi.org/10.1016/j.pragma.2003.10.012.
28. Sharar H, Al-Zagha AW, Ali E. ChatGPT in translating cultural nuances in O Night by Gibran Khalil Gibran. Int J Linguistics Literature Translation. 2025;8(3):30–8. https://doi.org/10.32996/ijllt.2025.6.4.6.
29. Shehab E, Badah AM. Translating conversational implicatures: A comparative study of human and artificial intelligence performance in Arabic–English literary translation. Translation Rev. 2026;124(1):38–54. https://doi.org/10.1080/07374836.2026.2617670.
30. Steen GJ. Analyzing metaphor in literature: With examples from William Wordsworth's I Wandered Lonely as a Cloud. Poetics Today. 1999;20(3):499–522.
31. Steen GJ, Dorst AG, Herrmann JB, Kaal AA, Krennmayr T, Pasma T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins. https://doi.org/10.1075/celcr.14
32. van den Broeck R. The limits of translatability exemplified by metaphor translation. Poetics Today. 1981;2(4):73–87. https://doi.org/10.2307/1772487.
33. 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

34. 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