

Decoding Figurative Language in Journalism Translation: Human and LLM Translation of English-Vietnamese Metaphors

Phuong Ba Luong*

Faculty of Foreign Languages, Academy of Journalism and Communication, Hanoi, Vietnam

Nguyen Thi Thanh Huong

English Department, Phenikaa University, Hanoi, Vietnam

Tran Quoc Viet

Department of English for Specific Purposes, Hanoi University, Hanoi, Vietnam

Abstract—The increasing use of large language models (LLMs) in professional translation has reshaped multilingual news production. While prior research has reported gains in fluency and surface-level accuracy, limited attention has been paid to LLMs' ability to translate figurative language, which plays a key role in meaning construction and evaluative stance in journalistic discourse. This study investigates the extent to which state-of-the-art LLMs achieve functional and cultural equivalence in the English–Vietnamese translation of journalistic metaphors, in comparison with professional human translators. Adopting a mixed-methods comparative design, the study analyzes a purposively selected corpus of 50 metaphorical sentences from major international news outlets. Translation outputs produced by two LLMs (GPT-4o and Claude 3.5) under two prompting conditions (zero-shot and advanced prompting) were compared with a human benchmark generated by senior Vietnamese news editors. All translations were evaluated through blind expert assessment based on four criteria: semantic accuracy, naturalness, cultural equivalence, and journalistic style. The findings show that advanced prompting significantly improves AI performance, particularly in semantic accuracy and stylistic fluency. However, LLMs continue to underperform in cultural equivalence, often exhibiting literalism and reduced metaphorical resonance. These findings underscore the indispensable role of human expertise as cultural and pragmatic mediators, suggesting that while LLMs accelerate production, human transcreation remains essential to preserving the rhetorical integrity of multilingual journalism.

Index Terms—large language models, metaphor translation, journalistic discourse, English–Vietnamese translation, cultural equivalence

I. INTRODUCTION

The rapid integration of large language models (LLMs) into professional translation workflows has profoundly transformed contemporary journalistic practices, particularly in multilingual news production. Advances in generative AI have enabled news organizations to accelerate cross-linguistic content circulation, reduce production costs, and respond to the increasing demand for real-time global reporting. As a result, translation in journalism is no longer a purely human activity but is increasingly characterized by forms of human–AI collaboration, in which automated systems generate initial drafts that are subsequently revised, evaluated, or curated by professional translators and editors.

Despite substantial progress in LLM-based translation with respect to fluency and surface-level grammatical accuracy, considerably less attention has been paid to the ability of these systems to handle figurative language. In journalistic discourse, figurative expressions, particularly metaphors, are not merely stylistic embellishments. Rather, they function as central framing devices that shape readers' interpretations, convey evaluative stance, and organize complex political, economic, and social realities. Inaccurate or inappropriate translation of such expressions may therefore lead to subtle but consequential distortions in meaning, tone, and ideological positioning in multilingual news dissemination.

The challenge of translating figurative language is especially pronounced in the English–Vietnamese language pair. English journalistic metaphors are often grounded in culturally entrenched conceptual domains such as WAR, FIRE, MOTION, and NATURAL DISASTERS, which reflect Anglophone media conventions and sociocultural assumptions. These metaphorical mappings do not always have direct or conventionalized equivalents in Vietnamese, a language with distinct typological features, discourse preferences, and culturally salient figurative patterns. Consequently, successful translation frequently requires strategies of adaptation, reformulation, or metaphorical substitution rather than direct lexical transfer. This places substantial cognitive and cultural demands on translators and raises critical

* Corresponding Author. Email: ltphuong2012@gmail.com

questions regarding the extent to which current LLMs can approximate such expertise.

Although a growing body of research has examined the performance of machine translation and LLM-based systems in general translation tasks, empirical studies focusing specifically on figurative language remain limited, particularly in domain-specific genres such as journalism and in under-researched language pairs such as English–Vietnamese. Existing studies tend to prioritize global accuracy metrics or sentence-level fluency, often underrepresenting higher-level pragmatic, cultural, and rhetorical dimensions of translation quality. As a result, there is insufficient empirical evidence regarding how LLMs handle metaphorical meaning in professional news translation and how their performance compares with that of experienced human translators.

Against this background, the present study investigates the translation of figurative language in English digital journalism into Vietnamese, comparing the performance of state-of-the-art large language models with professional human translators. Specifically, the study aims (1) to compare human and LLM performance in translating English journalistic metaphors into Vietnamese; (2) to examine strengths and weaknesses in LLM output across multiple dimensions of translation quality, including semantic accuracy, naturalness, cultural equivalence, and journalistic style; and (3) to identify recurrent types of metaphor-related translation errors and their underlying linguistic and conceptual causes. By addressing these objectives, the study seeks to contribute to applied linguistics and translation studies by offering empirically grounded insights into the current capabilities and limitations of LLMs in handling figurative language, while also informing responsible models of human–AI collaboration in multilingual journalistic practice.

II. LITERATURE REVIEW

A. Previous Studies on Figurative Language and AI Translation

Research on machine translation (MT) and, more recently, large language models (LLMs) has predominantly focused on general translation quality, with an emphasis on lexical accuracy and overall fluency (Castilho et al., 2022). However, empirical evidence suggests that figurative language remains a persistent challenge. Studies examining idiom and metaphor translation in MT systems have reported systematic tendencies toward literalism and inappropriate image transfer (Moorkens & O'Brien, 2023). Recent investigations into LLMs such as GPT-4 indicate that while they exhibit greater contextual sensitivity, they often struggle with culturally embedded expressions, resulting in translations that are fluent yet stylistically flattened (Kocmi et al., 2023). Recent evaluations of large language models have further shown that while they excel in grammatical accuracy, they still struggle with the cultural and cognitive nuances of figurative language (Kojima et al., 2023). Furthermore, modern frameworks for translation quality assessment now emphasize functional equivalence and communicative effect, particularly when dealing with complex linguistic structures in low-resource language pairs (Hendy et al., 2023). Notably, existing work has largely concentrated on high-resource language pairs, leaving the English–Vietnamese pair relatively under-investigated.

B. Research Gap

Despite increasing scholarly interest in LLMs, several gaps remain in the existing literature. First, few studies have systematically examined the translation of figurative language using multidimensional evaluation frameworks that capture semantic, pragmatic, cultural, and stylistic dimensions simultaneously. As noted by Lucas (2023), the integration of AI in newsrooms has fundamentally shifted the translator's role toward post-editing; however, there is still a lack of empirical evidence on how these models maintain metaphorical integrity under such professional constraints. Second, research on LLM performance often lacks comparison with professional human benchmarks grounded in real-world newsroom practice. Third, there is a significant lack of empirical data regarding how LLMs handle the typological and cultural distance inherent in English–Vietnamese metaphor translation. These gaps necessitate a context-sensitive analysis of AI capabilities in professional journalism.

III. THEORETICAL GROUNDS

The present study is grounded in an integrated framework that combines Cognitive Metaphor Theory and Functionalist Translation Theory. From a cognitive perspective, metaphors are viewed as powerful framing mechanisms that shape how readers interpret complex events (Lakoff & Johnson, 1980). This is complemented by Nida's (1964) concept of functional equivalence and House's (2015) functional-pragmatic model, which provide the criteria for evaluating translation success. In this study, a successful translation is not defined by formal correspondence but by its ability to fulfill a similar communicative function, preserving the evaluative stance and rhetorical force of the source text within the target language's discourse norms.

IV. RESEARCH METHODOLOGY

A. Research Design

This study adopts a mixed-methods comparative research design to examine differences between professional human translators and large language models (LLMs) in the translation of figurative language in English–Vietnamese digital

journalism. Quantitative analysis is used to compare translation quality scores across predefined evaluation dimensions, while qualitative analysis is employed to identify recurrent patterns of figurative translation strategies and errors. This design enables both statistical comparison and in-depth interpretive analysis, in line with applied linguistics approaches that emphasize triangulation between numerical and discourse-analytic evidence (Saldanha & O'Brien, 2014).

The study is evaluative and comparative in nature. Professional human translations are used as a reference benchmark representing current newsroom standards, against which LLM outputs are compared. This benchmarking approach is not intended to treat human translation as an infallible gold standard, but rather as an ecologically valid point of comparison reflecting professional norms and editorial expectations in Vietnamese journalistic practice.

Two experimental factors were manipulated: (a) translation agent (human vs. LLM) and (b) prompting condition for LLMs (basic zero-shot prompting vs. advanced Chain-of-Thought prompting). This design allows for examination of both overall human–AI differences and the specific contribution of advanced prompting strategies to figurative language translation quality.

B. Research Data and Corpus

The dataset consists of a purposively constructed corpus of 50 English sentences containing metaphorical or idiomatic expressions, extracted from major international English-language news outlets, including *The Economist*, *The Financial Times*, and *The New York Times*. These outlets were selected due to their frequent use of metaphorical framing and rhetorically sophisticated language, making them appropriate sources for testing high-level translation competence in journalistic discourse.

Corpus construction followed a multi-stage purposive sampling procedure. In the first stage, two independent linguists with expertise in discourse analysis and translation studies identified an initial pool of approximately 100 candidate sentences containing potential figurative language, following the linguistic metaphor identification principles established by Steen et al. (2010). In the second stage, these candidates were jointly reviewed and filtered based on two predefined criteria:

- Semantic opacity, defined as the extent to which the figurative meaning cannot be reliably inferred from the literal meanings of the component words.
- Cultural loading, defined as the degree to which the metaphor relies on culturally specific conceptual schemas commonly found in Western political, economic, or social discourse.

From this pool, 50 items were selected for inclusion in the final dataset. The corpus was balanced across five thematic domains (10 items each): Economics, Politics, Technology, Environment, and General Idioms. This thematic stratification was designed to ensure coverage of both relatively conventionalized metaphors and highly context-sensitive expressions, thereby enhancing the analytical breadth and representativeness of the dataset.

C. Human and LLM Translation Procedures

For each of the 50 source sentences, three translation outputs were generated: one professional human benchmark translation and two LLM translations under different prompting conditions. This resulted in a total of 150 translation outputs for evaluation (50 human, 50 LLM basic prompting, and 50 LLM advanced prompting).

D. Human Translation Benchmark

To establish a professional reference benchmark, two senior Vietnamese news translators/editors, each with more than 15 years of professional experience in international journalism and English–Vietnamese translation, independently translated all 50 source sentences. A third senior editor then reviewed both versions and selected the translation judged to be most natural, rhetorically effective, and appropriate for publication in Vietnamese news media. This three-stage procedure was designed to approximate professional newsroom workflows and to ensure that the benchmark translations reflected both linguistic accuracy and journalistic stylistic norms.

E. LLM Translation Protocol

Translations were generated using two state-of-the-art large language models: GPT-4o and Claude 3.5. The choice of GPT-4o and Claude 3.5 as the primary LLMs for this study was informed by their state-of-the-art performance in current linguistic benchmarks. According to the LMSYS Chatbot Arena and recent industry evaluations (2024–2025), these models consistently demonstrate superior reasoning and creative writing capabilities compared to their predecessors. Specifically, Claude 3.5 was selected for its reported strengths in nuance-heavy tasks, while GPT-4o was chosen for its robust cross-lingual capabilities, making them the most relevant candidates for investigating the complexities of figurative language translation. To enhance replicability and procedural transparency, a standardized interaction protocol was implemented for all LLM queries.

For each source sentence, two prompting conditions were applied:

- Basic zero-shot prompting, using the minimal instruction: “Translate the following English sentence into Vietnamese.”
- Advanced Chain-of-Thought prompting (Wei et al., 2022), using an engineered prompt that required the model to: (a) identify the figurative expression, (b) explain its intended meaning and pragmatic function in context, and (c) produce a final Vietnamese translation prioritizing functional and cultural equivalence.

All prompts were administered in isolated sessions, and conversational memory was reset between queries to minimize cross-item influence. Temperature and other stochastic generation parameters were held constant across all queries to reduce variability attributable to generation settings. Only the final Vietnamese translations were retained for evaluation; intermediate reasoning outputs were excluded to ensure blindness in the rating procedure.

F. Evaluation Criteria and Rating Procedure

All 150 translations were evaluated using four predefined criteria:

- Semantic Accuracy (SA): The degree to which both propositional and figurative meanings of the source text are correctly conveyed.
- Naturalness (NAT): The extent to which the translation reads fluently and idiomatically in Vietnamese.
- Cultural Equivalence (CE): The degree to which the figurative expression is rendered in a culturally appropriate and conceptually meaningful way for Vietnamese readers.
- Journalistic Style (JS): The extent to which the translation conforms to Vietnamese news discourse norms in terms of register, tone, and stylistic conventions.

Each criterion was rated on a 5-point Likert scale (1 = very poor, 5 = excellent). Detailed descriptors and illustrative examples were provided to guide rater judgment and enhance rating consistency.

G. Raters and Inter-Rater Reliability

To ensure a rigorous evaluation, we recruited three independent experts, all of whom hold doctorates in applied linguistics or translation studies and possess extensive experience in quality assessment. Before the formal rating began, these experts underwent a structured calibration session to align their perspectives. During this time, they thoroughly reviewed the four evaluation criteria and jointly graded a pilot set of 10 translations (excluded from the main study).

This process was essential to harmonize their interpretation of the scales and minimize subjective bias. The consistency of their judgments was then verified using Cronbach's alpha. The analysis revealed strong reliability across all criteria: semantic accuracy ($\alpha = .88$), naturalness ($\alpha = .86$), cultural equivalence ($\alpha = .91$), and journalistic style ($\alpha = .84$). Since these figures comfortably exceed the standard thresholds for applied linguistics research, they confirm that the rating procedure was both stable and robust.

H. Data Analysis

Quantitative analysis was conducted using descriptive and inferential statistics. Mean scores and standard deviations were calculated for each translation condition and evaluation criterion. One-way ANOVA tests were performed to examine statistically significant differences among the three translation conditions (human benchmark, LLM basic prompting, and LLM advanced prompting). Where appropriate, post-hoc comparisons were conducted to identify specific pairwise differences. Effect sizes were also calculated to provide an indication of the magnitude of observed differences.

To complement the quantitative findings, a qualitative error analysis was conducted focusing on AI-generated translations that received a mean score below 3.0 on cultural equivalence or naturalness. These outputs were subjected to inductive coding to identify recurrent patterns of figurative translation failure. Through iterative coding and rater discussion, three dominant error categories were established: (a) literalism, (b) collocational clash, and (c) rhetorical dilution. Representative examples were selected to illustrate typical patterns and to support interpretation of the quantitative findings.

I. Ethical Considerations

The study did not involve the collection of personal or sensitive data. All human translators and raters participated voluntarily and provided informed consent for their contributions to be used for research purposes. No personally identifiable information was collected or reported. The news texts used in the corpus were publicly available materials used for research and educational purposes. The study was conducted in accordance with standard ethical guidelines for applied linguistics and translation studies research, with particular attention to transparency, replicability, and responsible reporting of AI system performance.

V. RESEARCH FINDINGS AND DISCUSSION

A. Research findings

(a). Overall Performance Across Translation Conditions

Table 1 presents the mean scores and standard deviations for the three translation conditions. The results indicate clear differences in overall translation quality across conditions. Translations produced under the basic prompting condition obtained the lowest mean score ($M = 2.85$, $SD = 0.92$), suggesting relatively weak and unstable performance. In contrast, advanced Chain-of-Thought prompting was associated with higher and more consistent scores ($M = 3.92$, $SD = 0.54$). The human benchmark achieved the highest mean score ($M = 4.75$, $SD = 0.28$), with notably low variability, reflecting consistently high-quality translations.

A one-way ANOVA indicates that differences among the three conditions are statistically significant ($p < .001$). Post hoc comparisons further suggest that advanced prompting significantly outperforms basic prompting, while the human benchmark remains significantly higher than both LLM conditions.

TABLE 1
MEAN SCORE ACROSS TRANSLATION CONDITIONS (N = 50)

Translation condition	M (mean score)	SD (standard deviation)
LLM (Basic prompting)	2.85	0.92
LLM (Advanced prompting)	3.92	0.54
Human benchmark	4.75	0.28

(b). Performance Across Evaluation Criteria

To further examine specific dimensions of translation quality, mean scores for the advanced LLM condition and the human benchmark across the four evaluation criteria are reported in Table 2. The smallest difference is observed in semantic accuracy (LLM: M = 4.50; Human: M = 4.90), indicating that LLMs generally succeed in transferring core propositional meaning. A larger gap is evident for naturalness (LLM: M = 3.80; Human: M = 4.80), suggesting that AI outputs are perceived as less idiomatic in Vietnamese.

The most substantial discrepancy is found in cultural equivalence (LLM: M = 3.20; Human: M = 4.75), representing a difference of 1.55 points. This result indicates persistent difficulty for LLMs in achieving culturally appropriate metaphorical transcreation. Differences in journalistic style are comparatively smaller (LLM: M = 4.20; Human: M = 4.65), suggesting that LLMs are generally able to approximate professional news register.

TABLE 2
MEAN SCORES BY EVALUATION CRITERION (ADVANCED AI VS. HUMAN)

Criterion	AI (M)	Human (M)
Semantic accuracy (SA)	4.50	4.90
Naturalness (NAT)	3.80	4.80
Cultural equivalence (CE)	3.20	4.75
Journalistic style (JS)	4.20	4.65

(c). Distribution of Metaphor-Related Error Types

Qualitative coding of LLM outputs identified three recurrent categories of metaphor-related errors. As shown in Table 3, literalism accounts for the majority of identified errors (65%). These cases involve word-for-word renderings that preserve surface form but fail to convey intended metaphorical meaning in Vietnamese. Collocational clashes account for 20% of errors and are characterized by grammatically acceptable but pragmatically unnatural combinations. The remaining 15% involve loss of rhetorical force, where over-explication or neutralization reduces the evaluative and persuasive impact of the metaphor.

TABLE 3
DISTRIBUTION OF METAPHOR-RELATED ERROR TYPES IN AI OUTPUTS

Error type	Percentage
Literalism	65%
Collocational clash	20%
Loss of rhetorical force	15%

(d). Inter-Rater Reliability

To assess the consistency of metaphor identification and strategy coding, inter-rater reliability was examined on a randomly selected subset of the dataset comprising approximately 50 items. Agreement between coders was found to be high, indicating that the coding scheme was applied consistently across raters. These results support the methodological robustness of the quantitative distributions reported above.

(e). Illustrative Examples of Metaphor Translation Patterns

To complement the quantitative distributions reported above, this section presents illustrative examples that exemplify typical patterns observed in human and LLM translations. These examples are representative of recurrent phenomena identified during qualitative coding and are intended to make concrete the types of metaphor-related successes and failures observed in the dataset.

TABLE 4
ILLUSTRATIVE EXAMPLES OF METAPHOR TRANSLATION PATTERNS

Source metaphor (English)	Human benchmark translation (Vietnamese)	Typical AI translation (Vietnamese)	Identified issue
"The economy is facing strong headwinds."	"Nền kinh tế đang đối mặt với những lực cản mạnh mẽ."	"Nền kinh tế đang đối mặt với những cơn gió mạnh."	Literalism (surface-form transfer)
"The proposal triggered a political firestorm."	"Đề xuất này đã làm bùng lên làn sóng chỉ trích dữ dội."	"Đề xuất này đã gây ra một cơn bão lửa chính trị."	Collocational clash
"Tech stocks took a beating."	"Cổ phiếu công nghệ lao dốc mạnh."	"Cổ phiếu công nghệ bị đánh đập."	Literalism
"The policy opened the floodgates to criticism."	"Chính sách này đã mở đường cho làn sóng chỉ trích ồ ạt."	"Chính sách này đã mở cổng lũ cho sự chỉ trích."	Rhetorical dilution / unnatural imagery
"The talks hit a brick wall."	"Các cuộc đàm phán rơi vào bế tắc."	"Các cuộc đàm phán đâm vào bức tường gạch."	Literalism

A closer examination of specific cases in Table 4 reveals the underlying nature of these translation failures. In the case of 'political firestorm', for instance, the LLM produced 'cơn bão lửa chính trị'. This is a clear instance of collocational clash; while 'firestorm' is a standard metaphorical source domain for intense criticism in English, its direct equivalent in Vietnamese ('bão lửa') is typically restricted to literal military or fire-related contexts. A more functionally equivalent rendering would be 'làn sóng chỉ trích dữ dội' (a fierce wave of criticism), which aligns with Vietnamese journalistic conventions. Similarly, the literal rendering of 'took a beating' as 'bị đánh đập' or 'hit a brick wall' as 'đâm vào bức tường gạch' demonstrates the model's struggle to move beyond word-level semantic mapping to conceptual-level transfer.

These illustrative cases directly reflect the dominant error categories reported in Table 3, particularly the prevalence of literalism and collocational mismatch. By contrast, human benchmark translations consistently employ conventionalized target-language expressions to preserve rhetorical force and genre appropriateness".

B. Discussion

(a). Overall Performance Patterns and Functional Differentiation Between Human and AI Translation

The present study set out to investigate not only whether state-of-the-art large language models (LLMs) can approximate professional human performance in translating journalistic metaphors from English into Vietnamese, but also how differences in performance are distributed across dimensions of translation quality. The results reported in Tables 1–3 reveal a differentiated performance profile in which LLMs demonstrate substantial gains in semantic accuracy and surface-level fluency, particularly under advanced prompting conditions, while human translators maintain a consistent advantage in cultural equivalence, rhetorical force, and journalistic appropriateness.

This pattern supports and extends recent empirical findings in LLM-assisted translation research, which suggest that contemporary LLMs are increasingly effective in producing linguistically well-formed and semantically plausible translations, yet remain limited in their ability to handle pragmatically and culturally embedded meaning (Moorkens & Reiter, 2023; Zhu et al., 2024). In the present study, prompting led to statistically and substantively meaningful improvements in semantic accuracy (Table 1), indicating that prompting strategies can partially compensate for weaknesses in default model behavior. However, these improvements did not translate into equivalent gains in cultural equivalence or metaphorical force (Table 2), underscoring a persistent gap between linguistic competence and culturally grounded interpretive competence.

From a theoretical perspective, this differentiated pattern reinforces the argument that translation quality cannot be reduced to a unidimensional construct. Instead, the findings align with multidimensional models of translation quality assessment, such as MQM (Lommel et al., 2014) and House's (2015) functional-pragmatic framework, which emphasize that accuracy, fluency, style, and cultural-pragmatic adequacy constitute partially independent dimensions. The present results empirically demonstrate that LLMs may achieve high scores on dimensions that are more closely tied to lexical-semantic mapping, while underperforming on dimensions that require culturally situated interpretation and rhetorical judgment.

(b). Literalization and Metaphorical Attenuation as Systematic AI Failure Patterns

A central qualitative finding of the study concerns the tendency of LLMs to default to literalization strategies when confronted with metaphorical expressions, even under advanced prompting. As illustrated in Table 3 and in the illustrative examples, a substantial proportion of AI-generated translations exhibited metaphorical attenuation, whereby the figurative force of the source expression was weakened or replaced by a more neutral, descriptive paraphrase.

This pattern is theoretically significant for several reasons. First, it suggests that LLMs, despite their extensive exposure to metaphorical language in training data, may prioritize propositional meaning over rhetorical framing. This aligns with prior work showing that neural systems often favor semantically transparent paraphrases at the expense of stylistic or pragmatic equivalence (Moorkens & O'Brien, 2023). Second, the attenuation of metaphorical force has important discourse consequences in journalistic contexts, where metaphors frequently function as framing devices that shape readers' interpretations and emotional responses to news events (Charteris-Black, 2011; Schäffner, 2004).

The observed literalization strategies can also be interpreted through the lens of cognitive metaphor theory. According to Lakoff and Johnson (1980) and Kövecses (2010), metaphors are not merely linguistic ornaments but reflect entrenched conceptual mappings. Successful metaphor translation therefore requires not only recognizing a metaphorical expression, but also identifying an appropriate target-language mapping that is culturally conventionalized. The present findings suggest that LLMs often fail at this second step, opting instead for semantically safe but rhetorically flattened renderings.

This tendency has important implications for the notion of functional equivalence (Nida, 1964). While AI outputs may preserve core propositional content, they frequently fail to reproduce the pragmatic and affective impact experienced by source-language readers. As a result, LLMs may achieve a form of semantic equivalence without achieving functional or communicative equivalence, particularly in figuratively dense journalistic texts.

(c). Differential Performance Across Metaphor Types and Source Domains

While the overall findings point to a general tendency toward literalization and metaphorical attenuation in AI output, a closer examination of the dataset reveals that these effects are not evenly distributed across metaphor types and source domains. Qualitative coding indicates that LLM performance varies systematically depending on whether metaphors are drawn from highly conventionalized source domains (e.g., MOTION, CONTAINER, FORCE) or from more culturally and discourse-specific domains (e.g., FIRE, WAR, SPORTS, NATURAL DISASTERS in political framing).

Metaphors grounded in highly conventionalized domains such as MOTION (e.g., “the economy is slowing down”) and CONTAINER (e.g., “fall into recession”) were more likely to be translated by LLMs in ways that approximated professional human renderings. In such cases, Vietnamese possesses relatively stable and conventionalized equivalents (e.g., chậm lại, rơi vào suy thoái), allowing LLMs to rely on high-frequency collocational patterns learned during training.

By contrast, metaphors drawn from evaluatively and rhetorically charged domains such as FIRE (“political firestorm”), WAR (“battle over reforms”), and SPORTS (“tech stocks took a beating”) were associated with a markedly higher incidence of literalization and collocational clash. These domains are particularly salient in English-language journalistic framing but do not always map straightforwardly onto culturally conventionalized Vietnamese metaphor systems. As a result, LLMs frequently produced surface-level transfers that were semantically transparent yet pragmatically and stylistically incongruent.

This pattern aligns with cognitive linguistic accounts emphasizing that metaphor translation requires sensitivity not only to conceptual mappings but also to the degree of conventionalization in the target culture (Kövecses, 2010; Schäffner, 2004). The present findings thus suggest that LLMs perform relatively well when metaphorical mappings are highly entrenched cross-linguistically, but struggle when mappings are culturally specific or genre-bound. This provides a more fine-grained account of LLM limitations than global claims about metaphor processing and highlights the importance of domain-sensitive evaluation in applied translation research.

(d). English–Vietnamese Specificity and Typological–Cultural Distance

The English–Vietnamese language pair represents a particularly demanding testing ground for metaphor translation, due to substantial typological and cultural differences. Vietnamese, as an isolating and tonal language, relies heavily on pragmatic particles, discourse conventions, and culturally conventionalized idioms, including four-syllable expressions, to convey evaluative stance and rhetorical emphasis.

The present findings indicate that LLMs struggle to consistently select culturally appropriate Vietnamese metaphorical equivalents, often producing translations that are semantically interpretable but stylistically incongruent with professional Vietnamese journalistic norms. This result supports Toury’s (2012) notion of translationese, whereby target-language norms are subordinated to source-language patterns, resulting in pragmatic unnaturalness.

Recent applied translation research has emphasized that such effects are particularly pronounced in typologically distant language pairs and in culturally dense genres such as journalism (Moorkens & O’Brien, 2023). By providing empirical evidence from an underrepresented language pair, the present study contributes to broadening the empirical base of AI translation research, which has to date focused predominantly on high-resource European languages. The findings therefore raise important questions about the generalizability of optimistic claims regarding LLM translation performance across diverse linguistic and cultural contexts.

(e). Applied Linguistics Perspective: Figurative Competence and Advanced Language Use

From an applied linguistics perspective, the findings highlight figurative competence as a critical dimension of advanced language use that remains challenging not only for second language (L2) learners but also for AI systems. Research on L2 metaphor processing has long demonstrated that even advanced learners often rely on literal or transfer-based strategies when encountering culturally specific metaphors, leading to reduced pragmatic appropriateness (Littlemore & Low, 2006; Kövecses, 2010).

The parallel between L2 learner difficulties and LLM failure patterns is theoretically instructive. It suggests that figurative competence is not merely a function of exposure quantity, but depends on the acquisition of culturally entrenched conceptual mappings and discourse conventions. In this sense, LLMs may be seen as exhibiting a form of

“advanced but non-native-like figurative competence,” characterized by high fluency and semantic coverage but reduced cultural and pragmatic attunement.

This analogy provides a novel Applied Linguistics framing for understanding AI language performance. It opens up new avenues for comparative research between human L2 learners and AI systems, particularly in relation to the development of figurative and pragmatic competence. Such a perspective also underscores the continued relevance of human expertise in contexts where language is used not merely to convey information, but to construct stance, evaluation, and cultural meaning.

(f). The Role of Prompting: Gains and Structural Limits

The results demonstrate that advanced prompting strategies, including Chain-of-Thought prompting, can significantly improve LLM performance in terms of semantic accuracy and overall coherence. As shown in Table 1, Level 3 prompting substantially narrowed the gap between AI and human translators on accuracy-related dimensions. This finding is consistent with recent research indicating that prompting can elicit more structured reasoning and improved task performance in LLMs (Wei et al., 2022; Kojima et al., 2023).

However, the limited impact of prompting on cultural equivalence and metaphorical force (Table 2) suggests that prompting primarily optimizes processes that are already within the model’s representational capacity. While prompting can encourage more careful analysis and reduce obvious errors, it cannot fully compensate for the absence of deeply entrenched, culturally specific conceptual mappings. This highlights a structural limitation of current LLM architectures in handling culturally grounded figurative language.

From a theoretical standpoint, this finding supports the argument that certain dimensions of translation quality are not easily reducible to explicit instructions or reasoning prompts. Instead, they depend on long-term immersion in specific linguistic and cultural communities, a form of situated competence that current AI systems do not possess.

(g). Methodological Implications for Evaluating Figurative Translation

The present findings also raise important methodological issues regarding the evaluation of metaphor translation quality. Automatic metrics such as BLEU and COMET, while useful for large-scale benchmarking, are limited in their capacity to capture functional equivalence, rhetorical force, and culturally grounded meaning (Rei et al., 2020; Reiter, 2023).

The use of expert human raters in this study therefore constitutes a key methodological strength. It enabled the analysis to capture subtle pragmatic and stylistic dimensions that remain invisible to automated evaluation. This supports recent calls in applied linguistics and translation studies for greater reliance on human-centered, multi-dimensional evaluation frameworks, particularly in research on creative and figurative language (Castilho et al., 2022; Moorkens & O’Brien, 2023).

More broadly, the findings suggest that research on AI translation quality must move beyond narrow accuracy-based metrics and incorporate discourse-level and culturally sensitive criteria if it is to provide a realistic assessment of system capabilities in professional contexts.

(h). Pedagogical Implications for Translator Education and EFL Contexts

Beyond newsroom practice, the findings have important implications for translator education and advanced EFL pedagogy. Figurative competence has long been recognized as a hallmark of advanced proficiency, yet it remains underrepresented in many curricula, which tend to prioritize grammatical accuracy and lexical expansion over metaphor awareness and rhetorical framing.

The systematic patterns observed in LLM output provide a novel pedagogical resource. LLM-generated translations, particularly those exhibiting literalization and metaphorical attenuation, may be used as diagnostic tools in translator training programs. By comparing LLM output with professional human benchmarks, students can be guided to identify subtle pragmatic and rhetorical mismatches and to develop greater sensitivity to culturally conventionalized metaphor use in Vietnamese journalistic discourse.

Such an approach aligns with recent pedagogical models that emphasize contrastive analysis, reflective post-editing, and critical engagement with AI tools as part of translator education. Rather than viewing AI as a threat to professional expertise, the present findings suggest that AI can be repurposed as a scaffold for developing higher-order figurative and pragmatic awareness.

In EFL contexts, the results also underscore the importance of explicitly teaching metaphorical framing and genre-specific metaphor use. Given the parallels between LLM and L2 learner difficulties, instructional interventions that target metaphor recognition, cross-cultural mapping, and genre-sensitive reformulation may benefit both human learners and professionals working with LLM-assisted translation systems.

(i). Micro-Level Discourse Effects and Framing Consequences

At a micro-discursive level, the attenuation or distortion of metaphorical framing observed in AI translations may have cumulative effects on how news events are interpreted by target-language audiences. Metaphors in journalism function not only as stylistic devices but also as framing mechanisms that structure evaluation, causality, and agency attribution.

For example, the replacement of a metaphor such as “political firestorm” with a literal or weakened equivalent reduces the sense of intensity, urgency, and conflict encoded in the original framing. Over repeated instances, such systematic attenuation may subtly reshape the evaluative tone of translated news discourse, leading to a less emotionally charged and less rhetorically dynamic representation of events.

This observation resonates with discourse-analytic research on framing and ideology, which emphasizes that small shifts in metaphorical choices can have disproportionate effects on audience interpretation (Charteris-Black, 2011). From this perspective, LLM-mediated metaphor translation is not merely a technical matter but a site of discursive intervention, with potential implications for how international news narratives are localized and consumed.

(j). Implications for Newsroom Practice and Human–LLM Collaboration

The results have direct implications for professional newsroom practice. The evidence supports a Human–LLM collaboration model in which LLMs are deployed primarily as support tools for initial drafting, rough translation, and preliminary metaphor identification, while human translators and editors retain responsibility for cultural adaptation, rhetorical refinement, and final quality control.

Such a division of labor allows news organizations to benefit from the efficiency and speed of LLM systems without sacrificing cultural and stylistic quality. Importantly, the findings suggest that reliance on LLM without sufficient human oversight may lead to systematic flattening of metaphorical framing, with potential consequences for audience interpretation and journalistic integrity.

From a professional training perspective, the results also highlight the need to equip translators and editors with skills for critically evaluating and post-editing LLM output, particularly in relation to figurative language and cultural resonance. This reinforces the role of Applied Linguistics in informing curriculum design for translator education in the age of AI.

(k). AI Governance, Ethics, and Discursive Risk

Beyond operational concerns, the findings raise broader ethical and governance-related issues. Metaphors play a central role in framing public discourse, shaping ideological positioning, and influencing audience interpretation (Charteris-Black, 2011). Systematic distortion or attenuation of metaphorical framing by AI systems therefore constitutes not merely a technical limitation, but a discursive risk.

Current AI governance frameworks tend to prioritize factual accuracy, data protection, and bias mitigation, while paying comparatively limited attention to rhetorical and discursive effects. The present findings suggest that figurative language should be recognized as a high-risk category in LLM-mediated communication, particularly in journalism and public information contexts.

From an AI governance perspective, this underscores the need for domain-sensitive risk assessment and the development of guidelines that explicitly address figurative and evaluative language. Such measures are necessary to ensure that the efficiency gains offered by LLM-based systems do not come at the expense of subtle but consequential shifts in public discourse framing.

(l). Theoretical Integration and Contribution

The findings contribute to theoretical debates in both translation studies and applied linguistics by empirically demonstrating the limits of current LLMs in culturally embedded figurative translation. The study supports multidimensional models of translation quality, reinforces cognitive and functional approaches to metaphor, and provides an applied linguistics framing that conceptualizes figurative competence as a form of culturally situated expertise that remains resistant to automation.

By situating English–Vietnamese journalistic translation within this broader theoretical landscape, the study advances understanding of how LLM-based systems interact with culturally grounded meaning and highlights the continuing centrality of human expertise in professional translation practice.

Despite the significant insights provided by this study, several limitations should be acknowledged. First, the corpus was restricted to 50 purposively selected metaphorical sentences. While this sample size might appear modest for broader statistical generalizations, it was strategically chosen to allow for a rigorous, expert-led qualitative error analysis. As noted by Saldanha and O’Brien (2014), in-depth interpretive research in translation studies often prioritizes the granularity of analysis over sheer data volume. This size enabled a meticulous examination of cognitive and cultural nuances that larger, automated datasets might overlook. Future research could expand this scope by utilizing larger parallel corpora or exploring a wider range of journalistic genres to further validate these preliminary findings.

VI. CONCLUSION

This study provided a multidimensional comparative analysis of metaphor translation in digital journalism, contrasting professional human editors with state-of-the-art LLMs. The empirical findings reveal a bifurcated landscape: while LLMs demonstrate remarkable proficiency in surface-level semantic integrity and grammatical fluency, they encounter significant barriers in navigating the nuances of figurative language. The persistent discrepancy between human and machine performance, particularly in cultural equivalence, underscores the limitations of pattern-based linguistic modeling in capturing the conceptual and ideological weight of metaphorical framing.

Qualitative evidence suggests that LLM-generated translations often exhibit a “literalist bias,” leading to collocational clashes and rhetorical dilution. By defaulting to literal renderings, LLMs risk stripping journalistic discourse of its evaluative stance and persuasive force. Although advanced prompting techniques such as Chain-of-Thought protocols can mitigate errors in semantic transfer and register appropriateness, they cannot yet substitute for the deep-seated cultural intuition and pragmatic judgment of professional human mediators. In the high-stakes environment of international news production, human expertise remains indispensable for ensuring that translated content resonates meaningfully within the target socio-cultural context.

Ultimately, this research contributes to applied linguistics by offering a grounded framework for evaluating the limits of LLM-based systems in handling non-literal meaning. As AI integration in journalism accelerates, these findings advocate for a human–AI collaboration model that prioritizes human oversight in the “transcreation” of figurative language. Future research should explore the evolving boundaries of LLM competence across more diverse language pairs and complex rhetorical genres, fostering a more nuanced understanding of the intersection between cognitive linguistics, professional translation practice, and generative artificial intelligence.

DECLARATION OF THE USE OF AI

The author would like to acknowledge the use of ChatGPT-4o and Claude 3.5 for linguistic refinement and grammatical editing of this manuscript to ensure clarity. Furthermore, the core translation data analyzed in the methodology were generated by ChatGPT-4o and Claude 3.5 as part of the research design. The author has reviewed and edited all AI-assisted outputs and takes full responsibility for the final content of the article.

REFERENCES

- [1] Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2022). *Machine translation quality assurance: From logic to linguistics*. Routledge.
- [2] Charteris-Black, J. (2011). *Politicians and rhetoric: The persuasive power of metaphor* (2nd ed.). Palgrave Macmillan. <https://doi.org/10.1057/9780230319141>
- [3] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, H., Labaka, G., & Awadalla, H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation*. arXiv. <https://arxiv.org/abs/2302.09210>
- [4] House, J. (2015). *Translation quality assessment: Past and present* (2nd ed.). Routledge. <https://doi.org/10.4324/9781315679662>
- [5] Kövecses, Z. (2010). *Metaphor: A practical introduction* (2nd ed.). Oxford University Press.
- [6] Kocmi, T., Elekes, R., Federmann, C., & Hermann, U. (2023). WMT23 metrics shared task results: LLMs are robust, but reference-based metrics are still needed. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)* (pp. 113–132). Association for Computational Linguistics.
- [7] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.
- [8] Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
- [9] Littlemore, J., & Low, G. (2006). *Figurative thinking and foreign language learning*. Palgrave Macmillan.
- [10] Lucas, M. (2023). *AI in journalism: The evolution of news production and translation*. Oxford University Press.
- [11] Moorkens, J., & O'Brien, S. (2023). *Translation and technology*. Routledge.
- [12] Nida, E. A. (1964). *Toward a science of translating*. E. J. Brill.
- [13] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics.
- [14] Saldanha, G., & O'Brien, S. (2014). *Research methodologies in translation studies*. Routledge.
- [15] Schäffner, C. (2004). Political discourse analysis from the point of view of translation studies. *Journal of Language and Politics*, 3(1), 117–150. <https://doi.org/10.1075/jlp.3.1.09sch>
- [16] Schäffner, C. (2017). Metaphor in translation. In K. Malmkjær (Ed.), *The Routledge handbook of translation studies and linguistics* (pp. 247–262). Routledge.
- [17] Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins.
- [18] Toury, G. (2012). *Descriptive translation studies—and beyond* (Revised ed.). John Benjamins.
- [19] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.



Phuong Ba Luong was born in Thanh Hoa (Vietnam) on 16th April 1968. He went to University of Foreign Languages, National University and graduated with a good degree in training English for teacher in 1994 and a good degree in Business Administration in 2001. He earned master degree in Applied Linguistics in 2003 from a joint training program between Vietnam National University and La Trobe University, Australia. Phuong was granted PhD degree in Comparative Linguistics in 2017 in Academy of Social Science, Vietnam.

He has been working as a teacher of English for 27 years in Faculty of foreign languages, Academy of Journalism and Communication. His main responsibilities are teaching English for different types of learners, both majored and non-majored. He has published 27 research articles in journals and proceedings of domestic and international conferences, published 5 reference (Four as chief author), conducted 6 research projects at university level and 1 at ministry level. His main research directions are contrastive linguistics and translation studies.

Dr. Luong is an active member Scientific Studies Board in AJC. He has been granted a certificate of merits by Prime Minister for his great efforts and achievements in education and training causes. Email address: ltphuong2012@gmail.com ORCID iD: <http://orcid.org/0009-0004-5194-6338>



Nguyen Thi Thanh Huong was born in Ha Noi (Vietnam) on 6th December 1963. She went to University of Foreign Languages, Vietnam National University, Ha Noi and graduated with an excellent degree in training Russian for teacher in 1985 and a good degree in English for teacher in 1994. She earned master's degree in English Language in 1999 in University of Foreign Languages, Vietnam National University, Ha Noi. Huong was granted PhD degree in Linguistics and Literature in 2004 in Vietnam National University, Ha Noi and was conferred Associate Professor title in Linguistics in 2017.

She has been working as a teacher of English for 32 years in Faculty of foreign languages, Academy of Journalism and Communication and in Faculty of English language, Phenikaa University. Her main responsibilities are teaching English for English-major students, both undergraduate and postgraduate levels. She has published 53 research articles in journals and proceedings of domestic and international conferences, conducted 11 research projects at university level. Her main research directions are theory of linguistics, contrastive linguistics, language of media, language and culture, and translation studies.

Assoc. Prof., Dr. Nguyen was an active member of Scientific Studies Board in AJC and is a member of Scientific Studies Board in Phenikaa School of Foreign Languages and Social Sciences, PU. She has been granted a certificate of merits by Prime Minister for her great efforts and achievements in education and training causes. Email address: huong.nguyenthithanh@phenikaa-uni.edu.vn



Tran Quoc Viet was born in Nam Dinh, Vietnam, on June 2, 1980. He received his B.A. degree in English from Luong The Vinh University, Vietnam, in 2008, and his Ph.D. degree in Linguistics from the University of Social Sciences and Humanities, Vietnam National University, Hanoi, in 2017.

He is currently a lecturer in the English for Specific Purposes Department at Hanoi University, Hanoi, Vietnam. Previously, he served as Dean of the Faculty of Foreign Languages at Hanoi Metropolitan University. He has published several research articles and textbooks for graduate and undergraduate students. His research interests include linguistics, terminology, and contrastive linguistics.

Dr. Tran Quoc Viet serves on the editorial board of the Scientific Journal - Hanoi Metropolitan University. Email: viettq@hanu.edu.vn (ORCID iD: 0009-0006-6213-5068).