A Comparative Study of Humans and AI on Nuanced Language Translation

Zitong Wang[1]

Received October 16, 2025
Accepted March 2, 2026
Electronic access March 31, 2026

Non-literal language such as jokes, song lyrics, and sarcastic remarks often loses its impact when translated across languages. With the emerging age of artificial intelligence and machine translation, the question arises of whether machines are capable of preserving this "punch", and if they are, how they compare to human translators. While previous research has generally emphasized human superiority in non-literal translation, much of that work was conducted before the rise of large language models. To address this gap, this study translated 78 English sentences into French across three categories of linguistic nuance: figurative language, slang, and sarcasm. Translations generated by humans, GPT-4, and DeepL were then compared using the Multidimensional Quality Metrics (MQM) framework. Statistical analyses revealed that GPT-4 achieved the highest overall accuracy in this sample, significantly outperforming both human translators and DeepL. Its advantage was strongest and most consistent in slang, moderate but still significant in sarcasm. Humans, on the other hand, maintained a slight edge in figurative language. These findings contribute to debates and research in translation studies, computational linguistics, and cognitive science by demonstrating that large language models can no longer be dismissed as inferior for nuanced translation within sentence-level translation, though challenges remain in tasks requiring pragmatic inference or deep cultural context.

Keywords: Machine translation, nuanced language, computational linguistics, translation quality, large language models.

Introduction

Human communication is rarely literal. Meaning is established through nuance, implication, tone, culture, or context that goes beyond the literal semantic meaning of a sentence. Such linguistic nuance conveys attitudes, feelings, humor, irony, identity, and social relations. Translating this requires preserving cultural appropriateness and the speaker's intent, not word-for-word substitution. Correct translation involves recognizing conventionalized meaning and selecting an appropriate target-language equivalent.

Building on this perspective, the present study examines how efficiency, accuracy, and error profiles differ between human-translated and AI-translated texts in multilingual translation tasks involving contextually or culturally nuanced phrases. By comparing translations produced by human translators, GPT-4, and DeepL, we investigated how different systems handle non-literal language and whether advances in large language models have altered previously observed performance gaps between humans and machines. In doing so, the analysis engages with competing theoretical accounts of translation quality, including data-exposure explanations grounded in large-scale statistical learning and accounts emphasizing deep pragmatic understanding, inference, and social cognition.

Understanding who excels requires understanding the neural and cognitive pathways involved in nuanced languages. Neuroimaging research demonstrates that non-literal language comprehension requires neural processes more complex than the literal type. It includes not only classical language regions, such as the left inferior frontal gyrus (IFG), but also those that are associated with theory of mind[1]. Studies show that the comprehension of figurative language relies heavily on regions of the left hemisphere associated with semantic and analytical processes. In a quantitative meta-analysis, Bohrn, Altmann, and Jacobs (2012) found that significant clusters of activation for metaphor conditions were mostly left-lateralized. Additionally, sarcasm processing was correlated with activations in midline structures, including the medial frontal gyrus (medFG) and anterior cingulate cortex (ACC), all key areas for theory of mind and inferring others' beliefs and intentions[2]. By contrast, AI systems process all languages by converting them into numerical representations through vector embeddings. They simulate fluency without real "understanding"[3].

Despite these limitations, machine translation (MT) systems have revolutionized global communication in the past decade. Platforms such as DeepL and OpenAI's GPT-4 are

[1] *Lower Canada College, Quebec, Canada*

© The National High School Journal of Science 2026

now capable of producing high-quality and fluent translations across dozens of languages. DeepL is a neural machine translation system that leverages large bilingual corpora and deep learning to produce accurate translations with high fluency[4]. It is particularly effective for literal translations and fast processing of large volumes of text. OpenAI's GPT-4 is a large language model trained on massive multilingual datasets across diverse domains. Unlike traditional MT systems, GPT-4 can generate context-aware translations that account for broader discourse, idiomatic expressions, and cultural nuance[5]. They have facilitated cross-border business, tourism, education, healthcare communication, and research access. They are especially valuable in low-resource environments where human translators are unavailable[6]. However, even advanced context-aware systems still mishandle nuanced languages. They may produce grammatically correct sentences that nonetheless distort tone or invert intended meaning[7]. Prior studies show that humans achieved high accuracy in nuanced translation. Human translators reached 94% accuracy when translating sarcastic lines of The Big Bang Theory tv show[8]. Errors in translation are not insignificant in real-life situations. Mistranslation can cause confusion, offense, and reputational and financial loss to companies[9].

While prior research has studied specific nuance types such as metaphor, idiom, or sarcasm, few studies have adopted a multi-category approach comparing human and AI translation performance[10–13]. This gap left an incomplete picture of where AI translation excels, where it fails, and how its performance compares to human bilinguals across multiple kinds of linguistic nuance. This study sought to address this gap by conducting a controlled comparison of human and AI-generated translations across three distinct types of linguistic nuance: figurative language, slang, and sarcasm. With the MQM framework, this research evaluated not only accuracy but also error types made. The study further aimed to identify where MT falls short when translating nuanced language and to what extent human expertise remains essential in preserving meaning beyond the literal.

We examined two types of machine translation systems that are currently being used widely. Neural Machine Translation (NMT) systems, like DeepL, are designed specifically to translate text between two languages. They rely on large bilingual datasets to align words and sentences accurately through neural mappings. This makes them highly effective for literal translations and fast processing of large volumes of text. Large Language Models (LLMs), like GPT-4, are trained on massive multilingual corpora across many domains. They can generate translations while considering broader context and cultural nuance. LLMs require substantial computational resources, but their versatility makes them well-suited for nuanced language tasks that go beyond straightforward word-for-word translation. Situating the analysis within these two systems would systematically identify where traditional NMT systems succeed or fall short and where LLMs offer advantages. This would give an understanding of the trade-offs between precision and context awareness in machine translation, and of which one should be prioritized under which context.

It is important to note that the analysis is limited to a set of curated English-French text passages and does not cover all languages or translation directions. This study focused on specific error types and severity using the MQM framework rather than broader linguistic evaluation metrics.

Based on prior findings in translation studies and cognitive science, we advanced several hypotheses. First, it was hypothesized that human translators would outperform both AI systems in overall translation accuracy, with a particularly strong advantage in translating sarcasm. Second, the study predicted that distinct error types would cluster with specific categories of linguistic nuance, reflecting differing cognitive and computational constraints. Finally, it was expected that sarcasm would remain especially challenging across both human and AI translations due to its reliance on pragmatic inference and theory of mind. By systematically comparing human translators, GPT-4, and DeepL using the MQM framework, this study evaluates the extent to which translation quality is better explained by large-scale data exposure or by deeper pragmatic understanding.

Methods

Source and collection

Sentences containing figurative language were extracted from the OPUS (Open Parallel Corpus) project, which consists largely of professionally translated materials such as published books and parliamentary proceedings. For this study, sentences were drawn specifically from the Books and Tanzil subcorpora[14]. Contemporary slang expressions were gathered from Generation Alpha pop music lyrics, including *Last Rizzday Night*, *Rizz Queen*, *Massive Meme Low Taper Fade*, and *Rizzler* by Jelly House. English-to-French translations of these expressions were obtained through crowd-sourced contributions on the Musixmatch platform. Sarcastic sentences were selected from the TransCasm corpus[15], a bilingual dataset consisting of 860 unique sarcastic English tweets paired with their corresponding French translations. From the sources mentioned, a total of 78 English sentences were selected and divided evenly across three nuance categories ($n = 26$ per category). To ensure sufficient statistical power, a priori power analysis was conducted assuming a large effect size (Cohen's $d = 0.8$), a significance level (α) of 0.05, and a desired power of 0.80. Cohen's d represents the standardized difference between group means, α indicates the probability of a Type I error (false positive), and power is the probability

© The National High School Journal of Science 2026

of correctly detecting a true effect. These parameters indicated that a minimum of 26 items per condition was required. Sentence selection was conducted entirely through manual inspection. Each sentence was read and annotated by the first author, a bilingual researcher familiar in English-French translation. Hereafter, this individual will be referred to as "the researcher." Only sentences containing a single and unambiguous instance of one of the three predefined nuance categories were retained. To ensure consistency and accuracy in the gradings, a subset of 18 sentence pairs (6 from each category) was independently re-evaluated by a certified English-French translator. Throughout this paper, "human translations" refers collectively to translations produced by non-AI sources.

Nuance categories

To assess the performance of human and AI translators in handling nuanced language, sentences were categorized into three distinct types of linguistic nuance: figurative language, slang, and sarcasm. Each category was selected for its cultural or contextual specificity, which makes direct translation especially challenging. Definitions of the three categories were drawn from previous linguistic literature.

A figurative language expression is defined as any instance where a speaker conveys meaning that differs from the literal interpretation of the words used[16]. This can include a wide range of rhetorical devices such as metaphors, similes, analogies, personification, and hyperbole, each of which relies on indirect or imaginative associations rather than explicit denotation. A slang expression is defined as "an ever-changing set of colloquial words and phrases that speakers use to establish or reinforce social identity or cohesiveness within a group or with a trend or fashion in society at large[17]." A sarcasm expression is defined as irony in which the speaker conveys the opposite of the literal meaning with the intent to mock, criticize, or convey disapproval[18].

AI Text Generation

Each sentence was entered into the translation systems individually without any context. To eliminate the influence of prior queries or user-specific memory, all translations were conducted in incognito mode in Google Chrome without any login credentials. Such setting enhances the comparability of outputs across systems and minimizes confounding variables related to session memory. For GPT-4, the prompt used was the following: "Translate the following sentence into French: [Sentence]". No additional instructions were provided. The output generated by each model was recorded without editing. For DeepL, each sentence was manually entered one by one into the GUI, with the target language set explicitly to French. All other settings were left untouched, and the outputs were recorded as displayed.

Annotation Procedure

To evaluate the quality of translations produced by both human and AI systems, this study uses a manual error analysis framework by the Multidimensional Quality Metrics (MQM)[19]. MQM provides a flexible and standardized set of error types for linguistic quality assessment, which enables consistent annotation. It has been validated in both academic and industry settings[20].

For this study, the error taxonomy was limited to categories most relevant to the challenges of nuanced language, particularly figurative expressions, idioms, culture-specific references, and stylistic choices.

All sentence pairs were directly evaluated by the researcher using the MQM error typology. Evaluation was done separately for every single sentence with each instance of translation error tagged with both its error type and a severity level. The definitions and severity scale follow MQM conventions.

Where multiple error types co-occurred in a sentence, each was annotated separately. In addition, to prevent annotation bias, the annotator was blinded to the origin of the translation (i.e., human, GPT-4, DeepL) during the error assessment process.

MQM 2.0 Dual Evaluation Scorecard: Raw and Linear Calibrated

Client	Project Name		Translator		Evaluator		
	Humans, Bilingual translation						
@ClientName	Source Language	Target Language		Total Word Count	Evaluation Word Count (EWC)	Content Type	
	Eng	Fre		3000	100	Figurative language	
Score Calculation Parameters	Maximum Score Value (MSV)	Passing Threshold Raw QS	Passing Threshold Calibrated QS		Reference Word Count (RWC)	Acceptable Penalty Points/RWC	
	100	93.3333333	80		150	10	
Evaluation Results	Quality Rating (QR)	Raw Quality Score (RQS)	Absolute Penalty Total (APT)	Per-Word Penalty Total (PWPT)	Normed Penalty Total (NPT)	Calibrated Quality Score (CQS)	Calibrated Quality Rating (CQR)
	Pass	99.00	1	6.01	1.50	97.00	Pass

General Comment

Note to the LQA evaluator: Please provide a detailed comment on quality of the evaluated sample.

Severity Penalty Multiplier	0	1	3	9				
Error Type	Neutral Errors Count	Minor Errors Count	Major Errors Count	Critical Errors Count	Error Type Penalty Total (ETPT)	Weighted ET Penalty Points	Normed ET Penalty Points	Error Type Weight
Terminology								
Wrong term					0	0	0.00	1
Accuracy								
Mistranslation					0	0	0.00	1
Overtranslation					0	0	0.00	1
Undertranslation	1				0	0	0.00	1
Addition		1			1	1	1.50	1
Omission					0	0	0.00	1
Style								
Unidiomatic style					0	0	0.00	1
Awkward style					0	0	0.00	1
Inconsistent style					0	0	0.00	1
Language register					0	0	0.00	1
Audience appropriateness								
Culture-specific reference					0	0	0.00	1
Grand Total	1	1	0	0	1	1	1.5	

Fig. 1 Multidimensional Quality Metrics Scoring Interface. Example scoring is provided.

In addition to the previous annotation work, the raw error count was recorded using the official MQM Scorecard. Calibrated scores were computed by assigning weighted penalties based on severity: minor errors were weighted with a factor of 1, major errors with a factor of 3, and critical errors with a factor of 9. Neutral observations were noted for qualitative

© The National High School Journal of Science 2026

Table 1 Examples of source sentences alongside their corresponding translations by a human, GPT-4, and DeepL.

Category	Source Sentence	Human Translation	GPT-4 Translation	DeepL Translation
Figurative Language	She quickly removed her hat, and during the whole scene that followed, held it against her breast, inside out, like a nest resting in the bend of her right arm.	Bien vite, elle enleva sa coiffure, et, durant toute la scène qui suivit, elle la tint contre sa poitrine, renversée comme un nid dans son bras droit replié.	Elle ôta rapidement son chapeau, et pendant toute la scène qui suivit, elle le tint contre sa poitrine, à l'envers, comme un nid reposant dans le creux de son bras droit.	Elle enleva rapidement son chapeau, et pendant toute la scène qui suivit, le tint contre sa poitrine, à l'envers, comme un nid reposant dans le pli de son bras droit.
Slang	There's a skibidi in my bed.	Il y a un skibidi dans mon lit.	Y'a un skibidi dans mon lit.	Il y a un skibidi dans mon lit.
Sarcasm	All I want is world tour and perfect grades is this too much to ask for?	Tout ce que je veux c'est un tour du monde et d'excellentes notes est-ce trop demander?	Tout ce que je veux, c'est un tour du monde et des notes parfaites, est-ce trop demander?	Tout ce que je veux, c'est un tour du monde et des notes parfaites est-ce trop demander?

purposes but did not contribute to the score. This scoring system yielded a single aggregate quality score out of 100 for each translation output. To ensure consistent score normalization while accounting for sentence length, the Reference Word Count (RWC) was set to 3/2 the length of the Evaluation Word Count (EWC). For example, if a translated sentence contained 20 words, its RWC was set to 30.

Data Analysis

All analyses were conducted in R (version 4.4.1). Prior to analysis, raw CSV files were cleaned by filling missing values downward for multi-error rows, trimming whitespace, and setting Severity as an ordered factor (Perfect $<$ Neutral $<$ Minor $<$ Major $<$ Critical) using tidyverse[21].

Validation against expert ratings employed paired t-tests to compare average scores assigned by the researcher and the expert translator for the same sentences, which assess whether systematic differences existed between raters. Because translation scores may not follow a bell-shaped normal distribution, Wilcoxon signed-rank tests were also conducted as a robustness check; this test compares relative rankings rather than exact numerical differences. Agreement in scoring, error categorization, and severity was assessed using Cohen's Kappa, which measures how consistently two raters agree beyond what would be expected by chance, with weights applied to reflect differences in severity levels. Unlike simple percentage agreement, Cohen's Kappa accounts for agreement that could occur by chance, providing a more reliable measure of inter-rater consistency. These agreement measures were computed using the irr (Inter-Rater Reliability) package in R[22]. Hereafter, when the researcher was the translator, that is referred to as "Researcher" while the certified translator is referred to as "Expert". Visualizations were generated with ggplot2 from tidyverse. To assess overall differences in translation quality across systems, a one-way ANOVA was performed with Calibrated Score as the dependent variable and Translator (Human, GPT-4, DeepL) as the independent variable using base R functions. ANOVA evaluates whether mean scores differ across multiple groups by comparing between-group variance to within-group variance, summarized by an F-statistic, which represents the ratio of between-group variance to within-group variance; larger F-values indicate greater separation between group means relative to variability within groups. When the ANOVA indicated a significant overall effect, post-hoc pairwise differences were examined using Tukey's HSD to control family-wise error rates using base R functions. To investigate whether the performance of each translator depended on the specific type of linguistic nuance, we performed a two-way ANOVA with Translator and Nuance Category as factors, followed by Tukey's HSD to identify specific pairwise differences. In addition to these multi-group comparisons,

© The National High School Journal of Science 2026

Table 2 Translation Error Categories and Definitions Based on the MQM.

Category	Error Type	Definition
Terminology	Wrong Term	Use of a term that a domain expert would not use, or that gives rise to a conceptual mismatch.
Accuracy	Mistranslation	Target content does not accurately represent the source content.
	Omission	Content present in the source is missing in the target.
	Addition	Content included in the target is not present in the source.
	Undertranslation	Target content is inappropriately less specific than the source content.
	Overtranslation	Target content is inappropriately more specific than the source content.
Style	Awkward Style	Style involving excessive wordiness or overly embedded clauses, often due to inappropriate retention of source text style in the target text.
	Unidiomatic Style	Text is grammatical but unnatural in expression.
	Inconsistent Style	Style varies inconsistently throughout the text.
	Language Register	Use of a level of formality higher or lower than required by specifications or general language conventions.
Audience Appropriateness	Culture-Specific Reference	Content inappropriately uses a culture-specific reference that will not be understandable to the intended audience.

a targeted contrast was conducted to examine overall differences between human-produced translations and AI-generated translations. For this analysis, GPT-4 and DeepL scores were pooled into a single AI group and compared against human translations using Welch's two-sample t-test, which does not assume equal variances between groups. To analyze differences in error-type distributions across translation systems, we conducted Chi-squared tests of independence, with Translator as one factor and MQM error category as the other. Because some error categories contained sparse counts, p-values were estimated using Monte Carlo simulation with 10,000 replicates rather than relying on asymptotic assumptions. When omnibus Chi-squared tests indicated significant associations, follow-up pairwise comparisons were performed between translators for each error type to identify the sources of the observed differences. In addition, relative prevalence of specific error categories was examined by comparing observed proportions across translators. These proportional comparisons were used to characterize qualitative patterns in error behavior. To examine explicitation tendencies, output length differences were analyzed at the sentence level. For each translation, the number of added words was computed as the difference between target and source sentence lengths. Mean added-word counts were calculated for each translator. Because identical source sentences were translated by multiple systems, there is potential for shared sentence-level variance across observations. To assess whether this non-independence could bias the reported comparisons, we conducted covariance, correlation, and multicollinearity diagnostics among all predictors. These checks were used to ensure that the statistical analyses were not driven by redundant or highly correlated explanatory variables.

© The National High School Journal of Science 2026

Table 3 Error Severity Levels Defined by MQM and Examples.

Severity Level	Definition	Example	Explanation
Neutral	an error that differs from a quality evaluator's preferential translation or that is flagged for the translator's attention but is an acceptable translation	"Tel est le simple plan de cette demeure où j'ai passé les jours les plus troublés mais aussi les plus heureux de ma vie – la maison d'où nous lancions nos aventures et où elles revenaient s'écraser comme des vagues contre un rocher nu."	*Rocher nu* does not capture *bare rock*, instead translating as *naked rock*, not something that would be used in English.
Minor	an error that does not seriously impede the usability, understandability, or reliability of the content for its intended purpose, but has a limited impact on, for example, accuracy, stylistic quality, consistency, fluency, clarity, or general appeal of the content	"Je ne suis plus seul dans cette pièce; une grande ombre agitée et amicale se déplace le long des murs et va et vient."	*Pièce* vs *chambre*. The latter is perhaps more formal and fits with the tone of the source sentence.
Major	an error that seriously affects the understandability, reliability, or usability of the content for its intended purpose or hinders the proper use of the product or service due to a significant loss or change in meaning or because the error appears in a highly visible or important part of the content	"Je ne suis plus seul dans cette pièce; une grande ombre agitée et amicale se déplace le long des murs et va et vient."	Uses "vitré" to replace "glazing". Seriously impedes understanding.
Critical	an error that renders the entire content unfit for purpose or poses the risk for serious physical, financial, or reputational harm	"Fanum est taxé dans la piscine"	Uses "Fanum est taxé" to replace "Fanum's tax". "Fanum's tax" is a way to refer to food theft between friends. Unfit for purpose.

© The National High School Journal of Science 2026

Results

We first assessed the reliability of the scoring, the calibrated MQM scores, error categories, and severity ratings assigned by the researcher were compared against those assigned by the expert (Table 4). Because both raters evaluated the same set of sentences, a paired t-test was used to compare their calibrated scores as it determines whether the average difference between two related measurements is statistically significant.

The test revealed no statistically significant differences between the researcher and expert ($t = -1.64$, $df = 53$, $p = 0.108$), with a mean difference of -1.875 points. On average, the researcher's scores were slightly lower, but the difference fell well within the 95% confidence interval, indicating that the two raters produced broadly consistent evaluations. A Cohen's κ of 0.386 indicated fair-to-moderate agreement beyond chance, a level considered acceptable[23]. Severity judgments showed higher alignment: 63% of annotations matched exactly, with a Wilcoxon signed-rank test indicating no systematic differences ($V = 60.5$, $p = 1.0$).

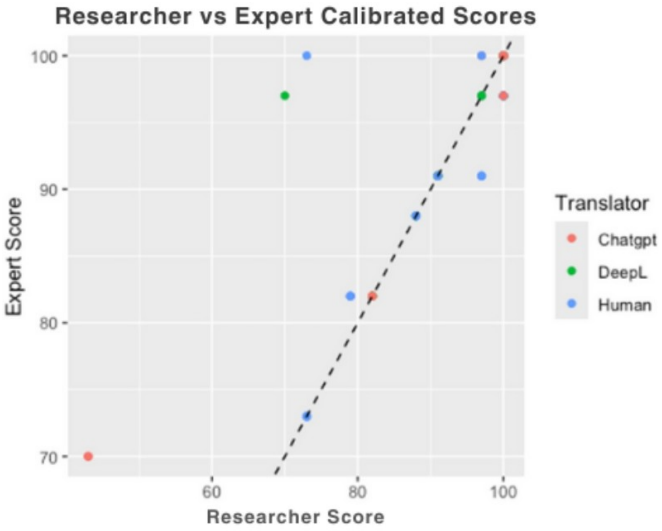

Fig. 2 Scatterplot comparing researcher-assigned scores with expert-assigned scores across three translation systems (GPT-4, DeepL, and Human). Each point represents a translation's evaluation with the dashed diagonal line indicating perfect agreement between researcher and expert scores. Deviations from the line reflect differences in evaluation alignment.

It was then tested whether translation quality, as measured by calibrated MQM scores, varied across systems. A one-way ANOVA revealed a significant effect of the translator on calibrated scores, with $F(2,231) = 6.92$, $p = .001$. Therefore, the between-system differences in mean translation quality were substantially larger than within-system variability. The mean scores followed a clear ranking: GPT-4 achieved the highest average calibrated score ($M = 91.0$), followed by human translators ($M = 85.7$), and DeepL ($M = 83.7$). The Tukey HSD test (Table 5) further reveals that DeepL and Human translators each produce significantly lower mean calibrated scores compared to GPT-4 (p-values of 0.0010 and 0.0240, respectively), with DeepL scoring 7.36 points lower and Human 5.34 points lower than GPT-4. However, there is no significant difference between Human and DeepL translators (p-value of 0.553), suggesting that GPT-4's superior performance was robust relative to both alternatives.

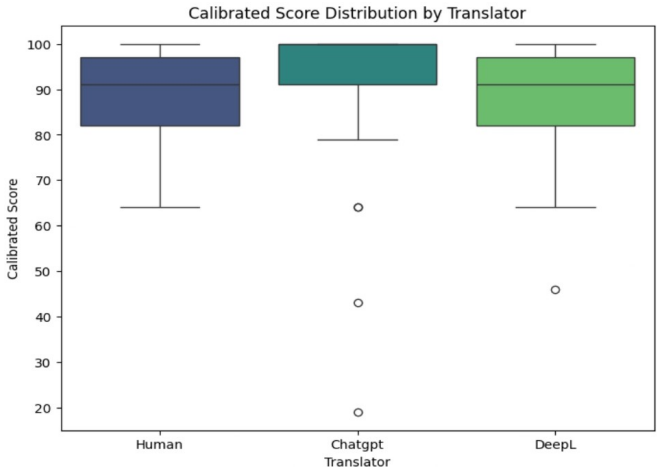

Fig. 3 Box plot illustrating the distribution of calibrated MQM scores for each translator (Human, GPT-4, and DeepL). Each box represents the interquartile range (IQR), with the line inside indicating the median.

However, the superiority of GPT-4 was not uniform across all nuance categories. The two-way ANOVA conducted with Translator and Nuance Category as factors revealed a significant interaction effect between the two factors ($F(4,262) = 2.80$, $p = 0.027$). Pairwise Tukey-adjusted post-hoc comparisons demonstrated that this interaction was driven primarily by slangs, where GPT-4 exhibited a distinct and statistically significant advantage. In this subset, GPT-4 scored significantly higher than both DeepL (mean difference $= 12.52$, $p = 0.002$) and human translators (mean difference $= 12.31$, $p = 0.002$). In figurative language, human translators achieved a slightly higher average score than GPT-4 (mean difference $= 3.60$) but did not reach statistical significance ($p > 0.05$). Moreover, sarcasm yielded the lowest absolute performance scores of all three categories for every translator.

To determine whether AI systems as a group outperformed humans, GPT-4 and DeepL were combined into a single "AI" category and compared it with human translators using a Welch two-sample t-test. This broader analysis did not reveal a statistically significant difference between AI ($M = 87.1$) and Human ($M = 85.7$), with t-statistic of 0.88 and a p-value

© The National High School Journal of Science 2026

Table 4 Inter-Rater Reliability and Agreement Between Researcher and Expert.

Measure	N	Statistic	95% CI / z	p-value	Interpretation
Cohen's κ (Calibrated Score, 54 subjects)	54	$\kappa = 0.386$	$z = 4.94$	7.68×10^{-7}	Fair-to-moderate agreement beyond chance
Cohen's κ (Severity, 54 subjects)	54	$\kappa = 0.478$	$z = 5.13$	2.96×10^{-7}	Moderate agreement beyond chance
Paired t-test (Calibrated Score)	54	$t = -1.64$, $df = 53$	$[-4.17,\ 0.42]$	0.108	No significant difference; student scores slightly lower

Table 5 Pairwise Comparisons of Mean Error Differences Between Translators.

Comparison	Difference in Mean	Adjusted p-value	Significance
DeepL – GPT-4	−7.36	0.0010	Significant
Human – GPT-4	−5.34	0.0240	Significant
Human – DeepL	2.02	0.553	Not significant

of 0.38, exceeding 0.05. Thus, while GPT-4 individually outperformed both Human and DeepL, the AI category as a whole was not consistently superior to human translators.

Beyond overall scores and error frequency, a chi-squared test of independence was conducted to examine whether translators differed in the types of errors they produced. The analysis revealed a significant association between translator and error category ($\chi^2 = 51.73$, $p = .0001$), indicating that error-type distributions varied systematically across translation systems. Because several cells exhibited low expected frequencies, p-values were computed using Monte Carlo simulation; accordingly, degrees of freedom are not reported[24]. Post hoc pairwise chi-squared tests showed significant differences between GPT-4 and DeepL ($\chi^2 = 32.15$, $p < .001$) and between GPT-4 and Human translations ($\chi^2 = 34.44$, $p < .001$). In contrast, no significant difference was observed between DeepL and Human translators ($\chi^2 = 1.33$, $p = .87$), suggesting comparable error-type distributions between these two conditions.

Furthermore, the error profile analysis showed that DeepL was associated with terminology-related errors, frequently defaulting to literal or domain-inappropriate lexical choices, particularly in figurative contexts. Human translations, while generally more stylistically natural, showed a higher incidence of accuracy-related errors, including mistranslations and omissions, especially when translating slang. In contrast, GPT-4 produced fewer accuracy errors overall but exhibited a relative tendency toward register mismatches and audience-inappropriate phrasing. Together, these results indicate that differences among translators extend beyond error frequency to variation in the kinds of errors they tend to make.

Table 6 Pairwise Post Hoc Comparisons.

Comparison	χ^2	p-value	Significance
GPT-4 vs. DeepL	32.15	$< .001$	Significant
GPT-4 vs. Human	34.44	$< .001$	Significant
DeepL vs. Human	1.33	.87	Not significant

We also compared the number of words added beyond the source sentence as differences in length are widely studied in translation research as indicators of explicitation where information that is implicit is made explicit[25]. On average, GPT-4 produced slightly longer outputs with 1.20 added words. DeepL with 0.77 added words and humans with 0.48 added words observed similar trends. However, given the overall sentence length of around 30 words, the additional words remain negligible.

To assess the robustness of the reported analyses, we examined the relationships among all predictors. Covariance and correlation analyses showed that Translator, Severity, and Nuance Category were only weakly to moderately associated,

© The National High School Journal of Science 2026

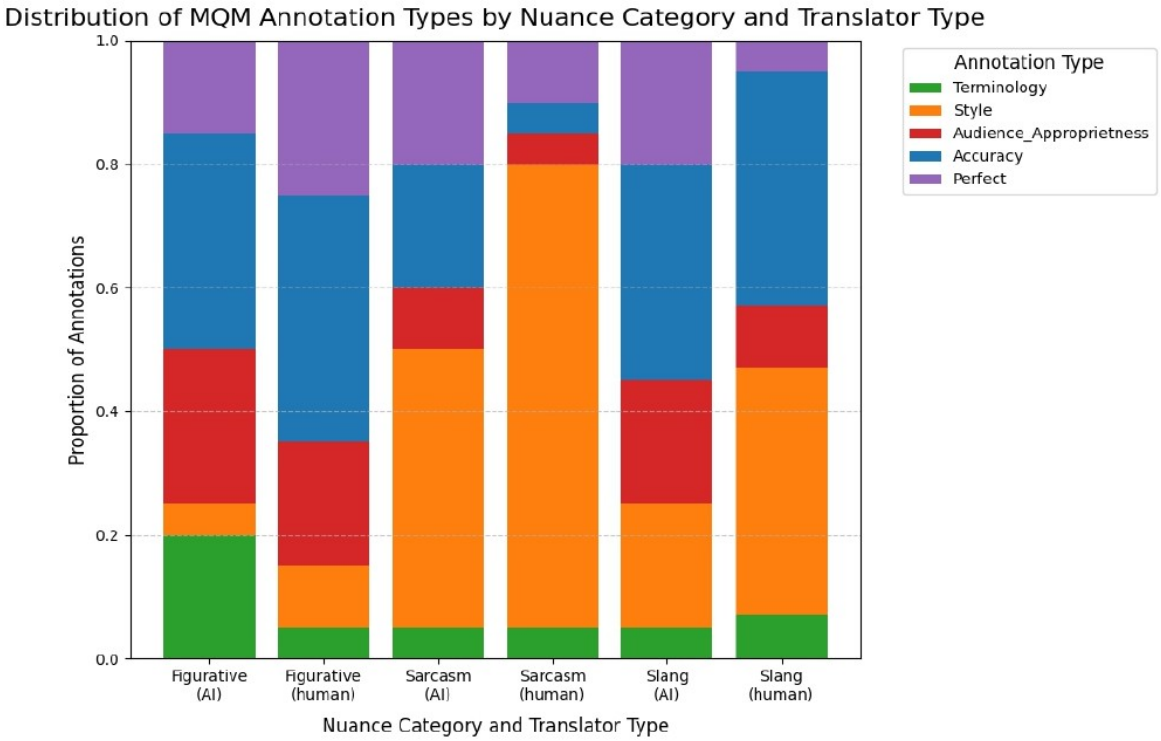

Fig. 4 Distribution of MQM error types by translator system (AI and human). The figure shows that AI translations contain a higher proportion of accurate and perfect annotations overall, particularly in slang and sarcasm, whereas human translations exhibit more issues related to style.

with all correlation coefficients below 0.36. Multicollinearity diagnostics further indicated minimal overlap among predictors, with all variance inflation factors below 1.06. Together, these results suggest that the predictors contribute largely independent information and that the statistical comparisons reported in this study are not driven by problematic dependence.

Discussion

This study compared the performance of GPT-4, human translators, and DeepL by using MQM scoring as the evaluation framework. Several consistent patterns emerged.

First, GPT-4 outperformed both human translators and DeepL in overall calibrated MQM scores and produced significantly fewer errors and higher mean quality. This advantage was important in slang and sarcasm where GPT-4 maintained stable performance while both human and DeepL translators struggled. These findings suggest that, at least in the context of nuanced language, GPT-4 demonstrates both higher average quality and greater reliability. The first hypothesis was therefore not supported.

Previous studies have generally found that GPT-4 performed on a similar level to junior translators but did not surpass experienced senior translators[26]. The results of this study partially align with this. For figurative language, human translators indeed performed better, which is consistent with prior findings. However, in the case of slang and sarcasm, GPT-4 matched or exceeded human performance. Furthermore, previous studies on sarcasm concluded that machine translation systems including GPT-4 often struggle with accurately translating sarcastic expressions. The results indicate GPT-4 slightly outperformed humans in this category but further highlight that sarcasm is a nuance category that is universally hard for both machine translation systems and human translators with minimal context. This supports the hypothesis that sarcasm is universally hard but contradicts the hypothesis that humans outperform GPT-4 in sarcasm. This conclusion, however, must be interpreted with caution due to the limited sample size. Moreover, Tang et al. (2024) demonstrated that GPT-4, when prompted appropriately, could generate high-quality, context-aware translations of East Asian idioms than human translators[27]. This study's findings differ in that GPT-4 did not outperform humans in figurative language; however, this could be due to the lack of context in our experimental design. GPT-4's strengths are not uniform across all nuanced categories.

© The National High School Journal of Science 2026

Humans' superior performance in translating figurative language might be explained through the lens of grounded cognition, which posits that conceptual understanding is deeply linked to sensory, motor, and social experiences. When encountering a metaphor, human translators do not rely solely on abstract linguistic rules; they simulate the scenario mentally[28]. For example, the phrase "kick the bucket" activates a mental model of death-related actions, emotions, and contexts. This embodied simulation allows humans to infer implied meaning, detect subtle cues, and preserve nuances in translation. These are capabilities that statistical models like GPT-4 or DeepL lack. Machine translation, even with enormous training data, does not "experience" the world and thus cannot internally simulate emotional, physical, or social contexts.

GPT-4's advantage in slang can be explained by its broad training exposure to internet forums and social media where slang is commonly found. These expressions are not hard to translate once their meaning is understood. Slang is especially dynamic and generationally marked, and even professional translators may struggle to keep up. By contrast, GPT-4's training on vast amounts of Internet text gives it wider exposure to these forms. It is also interesting to note that slang is highly group specific. Its meaning often relies on shared cultural and social knowledge, so translators who are not part of the relevant in-group cannot understand. These findings also have practical implications. GPT-4's strength in slang may suggest that AI translation systems could play a role in bridging generational gaps in language use. Younger speakers increasingly employ slang and sarcasm as important parts of online communication. This poses challenges for older speakers or professionals working across linguistic and cultural contexts. In such cases, AI tools may improve accessibility and understanding.

Unlike slang, sarcasm doesn't rely on unusual words or phrases. Instead, the words themselves are ordinary, but the speaker means the opposite of what they say. Understanding sarcasm therefore requires inferring intentions and attitudes that are not directly stated. LLMs and NMTs, trained on text rather than situated interaction, can only approximate these inferences statistically. Human translators, by contrast, integrate real-world knowledge and social cognition to resolve ironic intent. That being said, GPT-4 can still outperform humans because it has seen billions of examples of sarcastic language online. It detects patterns and linguistic cues associated with sarcasm, allowing it to approximate intended meaning, even without real-world social understanding. In isolated sentences, this exposure can give GPT-4 an edge over humans, who may misinterpret subtle cues without broader context.

As for the second hypothesis, accuracy errors were mostly associated with figurative language and slang while style errors, with sarcasm. This supports the hypothesis that different error types would cluster with specific nuance categories. The analysis of error types further revealed systematic differences in how systems fail and clarified each systems' profiles. DeepL leaned toward terminology errors, which reflects its domain-optimized neural machine translation architecture. GPT-4 was penalized more heavily for style and audience appropriateness. This shows that while it could capture meaning, its phrasing sometimes is not appropriate. These contrasts reflect architectural differences between NMTs and LLMs. Human translators were penalized more frequently for accuracy-related and style slips, which suggests that while they handled figurative language mostly correctly, they occasionally misread slang or sarcastic intent. Figurative and sarcastic categories showed wider dispersions in variance tests. This suggests inconsistency remains an issue regardless of system.

Interestingly, while GPT-4 consistently outperformed individual baselines, the combined "AI vs. human" comparison did not show a significant difference. This reflects the unevenness of AI systems. It is also worth noting that for this study, a non-personalized version of GPT-4 was used, with no user sign-in, as described under Methods; it is possible that the translation abilities of the AI differ depending on user-introduced biases; future work examining the degree to which the generic and user-tailored GPT-4 provide similar translation would provide important insight into the translation capabilities of artificial machines.

For the last hypothesis, the analysis showed that data exposure must be complemented by deep pragmatic understanding in order to produce coherent and high-quality translations. With one missing, translations may lack accuracy or cultural context. GPT-4 performed better than DeepL and humans in slang because it had access to that huge amount of training data. Despite this access, it still fails to generalize in certain cases. Figurative language requires abstract mapping between domains. This is a skill supported by human conceptual blending capacities that AI may struggle to replicate, which explains human superiority in this category, but they still failed in certain context because certain figurative expressions are less commonly used. Sarcasm is dispersed in all systems, and exposure to data alone cannot give accurate translation as explained previously.

Several limitations must be acknowledged. First, translations were evaluated at the sentence level rather than in extended discourse. Real-world communication often relies on paragraph or text-level context to clarify irony, tone, or figurative meaning, and this may have disadvantaged both humans and machines. Second, the corpora used were genre-specific: slang was drawn from music lyrics, figurative language from literary and religious texts, and sarcasm from online commentary. These choices do not capture the full range of how nuance appears in everyday communication. Third, although MQM provided a systematic evaluation framework,

© The National High School Journal of Science 2026

annotation involves subjective judgment, and some error categories inevitably overlap. Fourth, slang translations are user-generated and community-moderated and therefore do not constitute a professional or gold-standard reference. This study does not assume expert-level accuracy for these translations. Rather, they were used as a source of naturally occurring translation data to enable system-level comparison with AI-generated outputs. As established in prior work, certain types of linguistic data can be reliably collected from web-based contributors for comparative evaluation purposes[29]. Accordingly, conclusions drawn from the slang condition are restricted to relative performance patterns between human-produced and AI-produced translations and should not be interpreted as claims about absolute translation quality. Finally, machine translation systems evolve rapidly. As of 2025, newer generations of large language models have begun to emerge; therefore, the present findings should be interpreted as time-specific rather than permanent conclusions.

Future research could address these gaps by testing translations in larger datasets and including discourse-level context, which may help reduce misinterpretation of sarcasm and figurative expressions. Statistical modeling using computational approaches such as mixed-effects models could help isolate the influence of translator type, nuance category, and context length on translation quality. Expanding to additional systems would reveal whether the slang and sarcasm advantage persists or changes as models grow more sophisticated. Another promising direction is to examine hybrid workflows such as machine translation post-editing. The dataset could be expanded to include more domains, or even more languages, especially low-resource ones, and future work could explore complementary metrics such as fluency or adequacy judgments from end-users.

Conclusion

In conclusion, this study highlights how the rise of large language models may potentially reshape the translation landscape, particularly in handling nuanced language in English and French. While GPT-4 outperformed both human translators and DeepL overall especially in slang and with a modest edge in sarcasm, humans continued to show strengths in figurative expressions. These findings complicate older assumptions of consistent human superiority and go on to show that exposure to massive amounts of digital texts can give AI an unexpected advantage in certain contexts. At the same time, the persistence of errors in sarcasm serves as a reminder that translation is more than word substitution. It is a process that requires judgment, context, and cultural sensitivity. For everyday communication, AI offers unprecedented speed and accessibility, but when nuance matters most, human expertise may still remain indispensable.

Acknowledgments

I would like to extend my deepest gratitude to Dr. Andra Geana for her guidance and insightful feedback throughout this project. I am equally appreciative of Cheryl Cook, whose assistance with expert rating in the MQM scoring was essential to ensuring the reliability of my analysis.

References

1 P. A. Della Rosa, E. Catricalà, M. Canini, G. Vigliocco and S. F. Cappa, *NeuroImage*, 2018, **175**, 449–459.
2 I. C. Bohrn, U. Altmann and A. M. Jacobs, *Neuropsychologia*, 2012, **50**, 2669–2683.
3 Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich and N. Pinto, *Experience Grounds Language*, arXiv preprint arXiv:2004.10151, 2020.
4 DeepL, *DeepL AI Platform: Translation, Voice & Agentic Workflows*, `http://deepl.com/`, 2026.
5 J. Achiam, S. Adler, S. Agarwal *et al.*, *GPT-4 Technical Report*, arXiv preprint arXiv:2303.08774, 2023.
6 P. Naveen, P. Trojovský, Overview and challenges of machine translation for contextually appropriate translations. Iscience, 2024, 27(10).
7 Y. Jia and S. Sun, *Perspectives*, 2023, **31**, 950–968.
8 D. Anggraini, M. R. Nababan and R. Santosa, *International Journal of Multicultural and Multireligious Understanding*, 2020, **7**, 391–400.
9 M. Muhammadjonova, *Modern Science and Research*, 2025, **4**, 1382–1390.
10 I. Zaitova, B. M. Abdullah, W. Xue, D. Klakow, B. Möbius and T. Avgustinova, *It's Not a Walk in the Park! Challenges of Idiom Translation in Speech-to-text Systems*, arXiv preprint arXiv:2506.02995, 2025.
11 V. Dankers, C. Lucas and I. Titov, *Can Transformer Be Too Compositional? Analysing Idiom Processing in Neural Machine Translation*, 2022.
12 Z. L. Chia, M. Ptaszynski, M. Karpinska, J. Eronen and F. Masui, *Natural Language Processing Journal*, 2024, **9**, 100106.
13 I. Chamali, *It's All Greek to Them: Challenges in Translating Greek Slang and Idioms via LLMs and NMT*, 2025.
14 J. Tiedemann, Baltic Journal of Modern Computing, 2016.
15 D. Simon, S. Castilho, P. Lohar and H. Afli, *TransCasm: A Bilingual Corpus of Sarcastic Tweets*, 2022.
16 B. Dancygier and E. Sweetser, *Figurative Language*, Cambridge University Press, 2014.
17 C. C. Eble, *Slang & Sociability: In-Group Language Among College Students*, 1973.
18 N. Zhu and Z. Wang, *Personality and Individual Differences*, 2020, **163**, 110035.
19 A. Lommel, H. Uszkoreit and A. Burchardt, *Tradumàtica*, 2014, 455–463.
20 P. Charalampidou and S. Gladkoff, *Application of an Industry Practical Human MT Output Quality Evaluation Metric in the EMT Classroom*, 2022.
21 H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester and M. Kuhn, *Journal of Open Source Software*, 2019, **4**, 1686.
22 M. Gamer, J. Lemon, M. M. Gamer *et al.*, *Package 'irr': Various Coefficients of Interrater Reliability and Agreement*, 2012.
23 M. L. McHugh, *Biochemia Medica*, 2012, **22**, 276–282.
24 A. C. Hope, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1968, **30**, 582–598.

© The National High School Journal of Science 2026

25 P. Heltai, *New Trends in Translation Studies: In Honour of Kinga Klaudy*, BPH, Budapest, 2005.
26 J. Yan, P. Yan, Y. Chen, J. Li, X. Zhu and Y. Zhang, *Benchmarking GPT-4 Against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels*, arXiv preprint arXiv:2411.13775, 2024.
27 K. Tang, P. Song, Y. Qin and X. Yan, *Creative and Context-Aware Translation of East Asian Idioms with GPT-4*, arXiv preprint arXiv:2410.00988, 2024.
28 L. W. Barsalou, *Behavioral and Brain Sciences*, 1999.
29 R. Snow, B. O'Connor, D. Jurafsky and A. Y. Ng, *Cheap and Fast—But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks*, 2008.

© The National High School Journal of Science 2026