AI in specialised translation: Terminology, quality and the evolving role of human translators

Pascual Cantos-Gómez
Universidad de Murcia (Spain)
pcantos@um.es

Abstract

Generative AI systems are reshaping the position of specialised translation by performing tasks usually associated with professional translators. This paper examines whether generative AI can handle specialised terminology at levels of expert accuracy, how its translation accuracy fares alongside that of the human professional, and how, instead of replacing, AI may transform the role of translators. Preliminary findings appear to show that while models such as GPT-4 achieve a high terminological accuracy and are approaching mid-range human translation quality, they still require human supervision for nuances in domain-specific terminology. Rather than rendering the human translators obsolete, these AI tools are far more likely to transform translation work into a collaborative workspace which uses the efficiency of AI alongside human professionals. The paper ends with the need for new and targeted types of professional training and ethics to allow for this human-AI model in specialised translation.

Keywords: generative AI, specialised translation, human-AI collaboration, translation quality metrics, terminological accuracy.

Resumen

La IA en la traducción especializada: terminología, calidad y evolución del papel de los traductores humanos

La IA generativa está redefiniendo el ámbito de la traducción especializada, al realizar tareas propias de los traductores profesionales. En este artículo se examina el nivel de precisión terminológica de la IA en comparación con la de traductores expertos, así como la calidad de las traducciones generadas por IA

frente a las realizadas por traductores. También se analiza si la IA llegará a sustituir a los traductores o, más bien, modificar su rol dentro del proceso traductológico. Los resultados preliminares indican que, aunque las traducciones generadas por modelos como GPT-4 alcanzan niveles de precisión y calidad comparables a los de un traductor con experiencia intermedia, todavía es necesario que un ser humano las revise para garantizar el uso correcto de la terminología especializada. Más que sustituir a los traductores, la IA está convirtiendo el proceso traductológico en un espacio colaborativo que combina la eficiencia tecnológica con la competencia profesional de los traductores. El artículo concluye destacando la necesidad de incluir nuevas competencias en la formación y la ética profesionales de los traductores, con el fin de integrar de manera eficaz este modelo de cooperación entre la IA y los profesionales de la traducción especializada.

Palabras clave: IA generativa, traducción especializada, colaboración humana-IA, métricas de calidad traductológica, precisión terminológica.

1. Introduction

Large language model (LLM) development is gaining momentum and could have a profound impact on both research and practice in translation studies. With current technological capabilities, advanced LLMs can produce translations that demonstrate significant levels of fluency and coherence, using extensive multilingual databases and the implicit linguistic rules learned during training (OpenAI, 2023). Consequently, the technological capability to create large volumes of fluent translations has raised urgent concerns about their applicability in specialised domains such as technical, biomedical, legal or scientific fields.

Probably, three of the most pressing issues with the use of AI in specialised translation are: 1) Can an LLM utilise terminology specific to a certain area or industry with the same degree of precision as an expert within that field? Most biomedical terms were successfully translated by LLMs (Nazi & Peng, 2024). However, LLMs are unable to address ambiguities and nuances of domain-specific knowledge or polysemous meanings; therefore, some level of human intervention will be required. 2) What is the quality of LLM output compared to that of human translators? Benchmarking studies (e.g., Yan et al., 2024) show that LLMs are producing output comparable to that of junior/mid-level human translators, although benchmarking studies also reveal discrepancies in the outputs from LLMs, including term consistency,

discourse organisation and pragmatic awareness. 3) Is AI replacing translators, or simply redefining their role? The available data suggest that AI will best serve for lower-risk, repetitive-type translations (Zhong et al., 2024). Therefore, human translators may focus on refining/supervising higher-stakes, more complex translations.

Translator education has started to incorporate training on the utilisation of AI tools for translation tasks. These tools are gaining significance across all facets of the industry and are being created at an accelerating pace. Numerous translators currently use generative AI technologies in their processes for assistance (Zaim et al., 2025); however, they ultimately bear responsibility for the final output. The evolving landscape of translation education exemplifies this phenomenon. Translator education today encompasses more than the traditional linguistic ability previously mandated for the job. It now requires expertise and proficiency in rapid design, quality assessment and the management and maintenance of extensive datasets relevant to specific domains and/or languages. Many professionals in the translation business are now developing and utilising hybrid solutions. These hybrid systems enable AI to generate translations that can subsequently be checked, validated or amended by human translators. This signifies a significant transformation in the function of the translator. Translators now serve as key mediators, rather than simply providing translations.

The article explores these questions through a historical overview of machine translation development: from early rule-based and statistical models to current neural and generative architectures. It integrates comparative linguistic analysis and translation quality assessment between humans and AI-generated output, with particular attention to specialised fields. The article also examines the evolving role of the human translator in AI-augmented professional settings and considers the pedagogical and professional implications of these shifts. The article concludes with reflections on future directions for research, training and practice in specialised translation.

2. Evolution of machine translation paradigms

2.1. From rule-based systems to neural and generative models

Over the last few decades, machine translation (MT) has undergone significant paradigm shifts, beginning with handcrafted, rule-based systems

and evolving into data-driven neural networks, and more recently, into large-scale generative models. From the earliest period of MT development (the 1950s through the 1970s), Rule-Based Machine Translation (RBMT) dominated the field. RBMT used bilingual dictionaries and manually created grammar rules to translate the source language into the target language. Although it was groundbreaking at the time, it suffered from being overly complex. The 1966 ALPAC Report in the U.S. showed the dismal results from previous research efforts in early MT experiments and resulted in a temporary halt in funding for further research (Hutchins & Somers, 1992). However, some success was achieved in translating small, repetitive texts. For example, an example-based method was used in Canada to translate weather forecasts by using pre-translated sentence pairs for repetitive meteorological reports (Koehn, 2010). Overall, the RBMT paradigm was laborious to maintain, and it failed to produce acceptable translations when encountering sentences that did not match the predefined rules.

A breakthrough came in the 1980s through the 2000s with the introduction of Statistical Machine Translation (SMT) as the second paradigm of MT. SMT eliminated the need for explicit linguistic rules and replaced them with corpus-driven algorithms. Large numbers of parallel corpora were utilised to learn the probability of mapping phrases and reordering models, resulting in more fluent and idiomatic translations than those generated by the rule-based paradigm. In 2006, Google Translate's introduction of SMT provided a global platform to showcase the potential of data-driven translation technology to millions of users worldwide (Koehn, 2010). Despite these advancements, SMT had apparent limitations; it often produced disjointed and/or unnatural phrasing due to its focus on maximizing the local probability of n-grams, and it may have difficulty translating rare words or idioms that were underrepresented in the training data. Fluency and long-range coherence remained significant issues (Papineni et al., 2002).

In the mid-2010s, Neural Machine Translation (NMT) revolutionised MT with its use of deep learning to resolve many of the shortcomings in SMT. The ability to understand long-range dependencies and generate fluent translated sentences were two of the key factors in the success of NMT models (initially recurrent networks and then transformers), which convert entire sentences into high-dimensional vector representations. By 2016, Google Translate replaced SMT with an end-to-end neural model, resulting in improved overall quality of translation and lower error rates in certain scenarios (Bahdanau et al., 2015; Vaswani et al., 2017); although it was

disputed, a Microsoft Research study in 2018 asserted that human parity in translating Chinese to English on a news translation task (Hassan et al., 2018).

In the early 2020s, generative pre-trained language models, such as OpenAI's GPT-3 and GPT-4, have emerged and caused a paradigmatic shift in MT. They have begun to blur the distinction between MT and AI for generating natural language. These models, trained on massive quantities of multilingual text, can perform MT amongst other tasks via prompt-based learning (OpenAI, 2023). With billions of parameters, they can capture the nuances of language and domain-specific knowledge; and produce fluent and contextually aware translations. Recent studies have shown that GPT-4 produces translations comparable to or better than those produced by specialised MT systems in various standard benchmarking environments (Kocmi & Federmann, 2023; Yan et al., 2024). As well as this flexibility in generative models allows one model to support many different languages and domains (abilities that were formerly supported by separate MT systems), these models also present new concerns, such as how to maintain source fidelity and control outputs, particularly since the goal of generative AI is to generate the most probable continuation of a given text.

2.2. Linguistic characteristics of AI vs. human translation outputs

In addition to the overall quality of a translation, there are unique aspects of how AI translates compared to human translators, particularly in specialised areas. The handling of terminology and the lexical choice are two of the most significant aspects. Large-scale AI models show an impressive ability to recognise and correctly translate terminology, which human translators may not even be aware of (Mohamed et al., 2024). For example, in a recent study on translating scientific texts, GPT-3.5 achieved a significantly higher accuracy rate in translating domain-specific terms than a group of junior translators in training (Yan et al., 2024). However, this broad knowledge base comes at the cost of terminological consistency. Human translators are trained to apply terminology uniformly and follow established client or industry glossaries, whereas a generative model may vary its lexical choices. For example, a medical term such as *acid reflux* would ideally always be translated to *reflujo ácido* in Spanish. However, a generative MT may occasionally replace it with a synonym, such as *ácido gástrico retrógrado*, thereby reducing terminological consistency (Yan et al., 2024; Briva-Iglesias et al., 2024).

Additionally, AI systems face challenges when translating neologisms, abbreviations or acronyms within a specific field due to data sparseness (insufficient data to reliably estimate probabilities) in the training data. Human specialists typically infer neologisms or consult domain-specific sources, whereas generative AI models may provide a literal or incorrect translation, or a hallucination (an output that appears plausible but is factually incorrect, fabricated, or unsupported by data). Such errors can pose dangers in areas such as biomedicine and law, further emphasising the need for humans to review all the output generated by generative models (Mohamed et al., 2024).

Differences also appear at the syntactic and discourse levels. Human translators often reorganise sentences for improved readability or to meet stylistic requirements of the target language; this requires both creative ability and a deep level of comprehension of context. For example, professional translators often convert sentences in the source language from passive voice to active voice or break down complex sentences into two for clarity and readability. Conversely, unconstrained AI models generally replicate source sentence structure more literally. A study comparing English-Chinese scientific translations found that translators produced a larger number of shorter sentences, actively rearranged the organisation of the information and used a greater number of active voice constructions. In contrast, ChatGPT's output mirrored the length and structure of the source sentences, including the use of more passive voice (Fu & Liu, 2024). Similarly, in legal documents, translators adjust the tone and style of the document (i.e., select formal register or legal formulae), which AI does not reliably do unless prompted (Altakhaineh et al., 2025).

A fundamental difference in translation philosophy underlies some of the differences described above. Generative models are trained to reproduce the input text with high fidelity; they have no inherent motivation to add or subtract information from the input that maximises likelihood relative to learned patterns. In contrast, humans translate with an interpretive philosophy, which may include reading between the lines and utilising prior knowledge. Consequently, humans may make subtle additions or clarifications to the source text that were only implied. In a recent study, GPT-4 correctly retained the intended sense in "entering his 2nd year", whereas a human translator misread it as referring to a toddler, to a two-year-old child (Yan et al., 2024).

While AI offers uniformity, rapid production and strict adherence to the source text, human translators provide a better sense of context, adaptability to cultural differences and means of controlling the quality of the translation. Comparative studies found that while neither strategy outperforms the other, each one excels in distinct ways (Briva-Iglesias et al., 2024; Yan et al., 2024). Accordingly, there is an increasing trend towards hybrid models for translation, as these models combine AI with human expertise to create higher-quality translations than either method alone (Fu & Liu, 2024).

3. Evaluation of translations: automatic metrics and human assessment

Translation quality assessment (both for human translators and for AI) is one of the central methodological challenges in translation studies. As MT has evolved from rule-based and statistical models to neural and generative models, evaluation methods have likewise expanded to include both quantitative accuracy and qualitative insight. Automatic metrics enable large-scale, consistent evaluation, while human assessments remain the gold standard, as they capture both the semantic equivalence and pragmatic nuances, as well as the stylistic appropriateness of the translation. Both approaches offer unique insights when assessing the quality of specialised translations, which require a high degree of terminological precision, coherence and stylistic consistency.

3.1. Automatic metrics: Formulations, strengths and limitations

Automated metrics aim to replicate human judgements using mathematical models to quantify translation quality. The most established ones include *BLEU, METEOR, TER, chrF*, and *COMET*; each of these is based on a slightly different view of what constitutes equivalent to a human translation.

BLEU (*Bilingual Evaluation Understudy*; Papineni et al., 2002) measures *n*-gram overlaps between the machine translation and the human reference, with a brevity penalty term to discourage short translations. Its formula is:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right)$$

where p_n represents the modified *n*-gram precision, w_n are weights (commonly $1/N$), and BP = $\min(1, e^{1-r/c})$ penalises translations in comparison to longer reference translations (r = reference length, and c = candidate length). The BLEU score can vary from 0 (no overlap) to 1 (perfect match). While widely used, BLEU favours repetition of exact wording over equivalent meaning and provides low correlation between its scores and human evaluation at the segment level. In addition, two correct human translations of the exact text typically achieve only 0.6–0.7 BLEU against each other (Vashee, 2019); this indicates that the BLEU metric penalises lexical variety in an otherwise acceptable translation. For example, if "The patient experienced severe headache" is the reference and the candidate translation is "The patient had a strong headache", then BLEU will reduce its score because, although they are semantically equivalent, the *n*-grams "patient experienced" and "patient had" differ. This illustrates BLEU's lexical bias**,** as it rewards literal matching over semantic equivalence.

METEOR (*Metric for Evaluation of Translation with Explicit ORdering*; Banerjee & Lavie, 2005) improves on BLEU by incorporating morphology and synonymy. It aligns candidate and reference words using stemming and synonym dictionaries, computing a harmonic mean of precision *P* and recall *R*:

$$\mathrm{F}_{\alpha} = \frac{PR}{\alpha P + (1-\alpha)R'}$$

where adjusts the weight of precision versus recall. A penalty of fragmentation, *Pen*, is applied to reflect word-order differences (disorder alignments). The overall score is then computed as follows:

$$METEOR = F_{\alpha}(1 - Pen)$$

Because it considers synonymy and morphological variation, METEOR correlates more strongly with human ratings, particularly in morphologically rich or terminologically constrained languages.

TER *(Translation Edit Rate*; Snover et al., 2006) counts the minimum number of edits (insertions, deletions, substitutions, shifts) needed to transform the candidate into the reference:

$$TER = \frac{E}{R'}$$

where *E* represents the number of edits and *R* the reference length. A lower TER implies higher quality, reflecting the post-editing effort. However, it penalises legitimate reformulations that may be semantically accurate.

chrF (Popović, 2015) evaluates translation quality at the character level, using character *n*-grams to compute F-scores:

$$chrF_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 P + R'}$$

where *P* and *R* stand for precision and recall of character *n*-grams and *b* weights recall. This technique is most suited for languages with complex morphology (e.g., Finnish, Turkish) or those that exhibit agglutinative morphology. It also performs much better than BLEU in cases where there are smaller datasets or specialised terminology.

More recently, neural-based metrics, such as BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020), have been introduced. Unlike BLEU, these techniques measure similarity based on semantic representation instead of surface form. COMET uses high-dimensional contextual embedding to compare the source text, hypothesis and reference. COMET then trains a regression model on the encoded vector representations of the three inputs to approximate expert judgment:

$$COMET(s, h, r) = f_\theta\big(Enc(s), Enc(h), Enc(r)\big)$$

where f_q is the trained score function and achieved correlation scores greater than 0.9 when compared to Multidimensional Quality Metrics (MQM) human ratings on benchmark evaluations, such as the Seventh Conference on Machine Translation WMT 2022 (https://statmt.org/wmt22; Freitag et al., 2022), significantly outperforming both BLEU (0.45) and TER (0.53). The comparable evaluation data of several commonly used evaluation metrics are summarised in Table 1. The data support the notion that although BLEU and TER are still helpful for benchmarking purposes, evaluation metrics that are semantically based, such as COMET, are much more closely aligned with human expertise, a critical need for specialised translation.

Metric	Correlation with human judgment (ρ)	Strengths	Limitations
BLEU	0.45	Simple, reproducible	Ignores synonymy; surface-level comparison
TER	0.53	Reflects post-editing effort	Penalises reformulations
METEOR	0.61	Accounts for synonymy and stemming	Alignment–sensitive
chrF	0.67	Handles morphology; language-independent	Less effective for long sentences
COMET-22	0.92	High semantic fidelity	Computationally demanding

Table 1. Relative performance of key automatic translation metrics (Freitag et al., 2022).

These newer metrics extend beyond word-by-word overlap to include semantic adequacy, fluency, and contextually appropriate wording, making them especially useful in specialised translation, where paraphrasing and conceptual accuracy are much more important than reproducing a source sentence literally. For example, when two acceptable translations each employ different, but equally correct terms, such as "adverse reaction" versus "side effect", BLEU scores are reduced. At the same time, COMET maintains a higher value because the embeddings represent similar meanings. This demonstrates that neural metrics model human comprehension more effectively than do surface-level metrics. The shift from lexical to semantic metrics reflects a broader transformation in MT towards deeper language understanding.

3.2. Human evaluation: Multidimensional Quality Metrics and beyond

Human evaluators remain the most effective method for determining whether a translation is both accurate and communicatively effective. Human evaluators can determine whether a translation effectively conveys its intended message, as well as assess quality and ensure it uses the correct terms. Traditional approaches for evaluating translations have included direct assessment (DA), pairwise ranking and post-editing effort; however, these methods are prone to variability in judgments between raters.

The Multidimensional Quality Metrics (MQM) system, developed by the QTLaunchPad Consortium (Lommel et al., 2014), aims to standardise human evaluation. MQM defines a hierarchical categorisation of types of errors in translations (accuracy, terminology, style, fluency, grammar, omissions and additions), and each type of error has a corresponding severity rating (minor, major or critical). In addition, MQM provides a means to rate specific errors, which offers both granular detail and the ability to reproduce ratings. A total score is calculated by deducting weighted penalties

from an initial perfect score of 100 or by calculating a normalised accuracy ratio:

$$MQM\ Score = 100 \sum_{i=1}^{n} w_i e_i$$

where e_i represents each error instance and w_i is its severity weight. For example, in a medical source text "The patient exhibits photophobia", translating "photophobia" as "fear of light" would be marked as a critical terminology error, whereas "light sensitivity" would be correct and unpenalised. Instead, automatic metrics will likely treat both translations equally well, as both have the same lexical content.

For reliability purposes, the expert MQM annotators in WMT 2021-2023 achieved an average κ of approximately 0.75-0.80 (Freitag et al., 2022), which is considerably more reliable than DA ($\kappa \approx 0.45$). This demonstrates that trained professionals using MQM produce reliable and understandable results. Additionally, unlike BLEU or COMET, MQM can provide diagnostic information. One can determine if there are specific, systematic failures, such as frequent misinterpretations of terms or translation of specific proper nouns.

Metrics like BLEU and COMET measure similarity between two strings of characters, while MQM assesses whether the translated text is adequate, accurate and complete. While BLEU rewards literalness, MQM penalises literalness when it is the wrong choice. COMET uses context embedding to predict quality overall but cannot explain why a particular translation failed. Although MQM takes longer to evaluate and is more expensive than either BLEU or COMET, MQM's ability to validate interpretation is essential in those fields where a single error (i.e. *benign* versus *malignant*) can change the intended meaning.

Therefore, the complementary relationship between metrics and MQM is epistemic: metrics provide the benefits of consistency and scalability, while MQM provide the benefits of validity and interpretability. Today's best evaluations typically use a combination of both metrics and MQM: first, automated metrics are used to screen out bad candidates, and then a second stage of MQM-based human review is applied to those candidates who pass the initial screening.

Although the evaluation metrics have improved to match human evaluations, they still do not fully align with human evaluations. BLEU has shown a

moderate correlation with both MQM and DA ($\rho \approx 0.40$-0.50). In contrast, neural metrics (i.e. COMET) have demonstrated a high correlation with human evaluations ($\rho \approx 0.90$). However, correlations do not imply agreement with human translations. Evaluators often find contextual or pragmatic differences in translations that metrics ignore. This became evident in Microsoft Research's claim of "human parity" in the Chinese-English translation system (Hassan et al., 2018), based on BLEU. Human evaluators later identified contextual inaccuracies and unnatural writing styles within the machine-translated content (Fischer & Läubli, 2020). Therefore, merely achieving high metrics alone is insufficient to guarantee equivalent communication quality in translations.

Automatic metrics are replicable but may lack external validity. Human evaluations, on the other hand, are subjective, yet provide insights into communicative realism. Therefore, reliability will depend on how much the evaluation is triangulated, that is, combining quantitative metrics and qualitative human analysis.

4. Comparative evaluation of AI and human translation in specialised domains

The rapid proliferation of LLMs, including GPT-4, as well as Google's and DeepL's advanced systems, has impacted specialised translation. Recent empirical studies across legal, biomedical, and scientific domains consistently demonstrate a partial degree of convergence between the quality of translations produced by humans and those by AI systems (Fu & Liu, 2024; Moneus & Sahari, 2024). Quantitatively, state-of-the-art LLMs can achieve at least equal, if not better, levels of lexical accuracy and consistency compared with human translators. However, human translators continue to excel in conveying contextual nuance, making pragmatic judgments, and adapting to stylistic preferences. Although AI systems can produce fluent and accurate sentence-by-sentence translations, they still fall short in accurately capturing tone, idiomatic expressions, and domain-specific contexts.

4.1. Performance comparison in specialised domains

Studies highlight different strengths of AI and human translation in the legal domain. Briva-Iglesias et al. (2024) evaluated English legal documents

translated into Spanish, French and German using Google Translate, DeepL and GPT-4. Google Translate achieved the highest BLEU Score (47.3), outperforming GPT-4 (42.5) in surface-level similarity. However, COMET, which assesses meaning, favoured GPT-4 (0.82 vs. Google = 0.78). Legal professionals rated GPT-4's adequacy (4.5/5) and terminological accuracy (~95%) higher than DeepL (4.2/5; ~91%) and Google (3.9/5; ~88%). The difference between the BLEU and expert ratings underscores that surface metrics may misrepresent translation quality, as GPT-4 more accurately conveyed meaning despite lower literal overlap.

A similar pattern emerged in the biomedical domain. Lu et al. (2025) compared GPT-4, GPT-3.5, Google Translate and human translators on translations of patient-reported outcomes from English into six different languages. Based on automatic evaluation metrics, GPT-4 performed best (METEOR average = 0.81), followed by GPT-3.5 (0.78), human translators (0.76), and Google (0.74). Furthermore, GPT-4 produced equivalent results to human translations in terms of statistical significance ($p > 0.05$) in all four out of six possible comparisons. These findings suggest that, in controlled environments, LLM-based systems have the potential to produce results comparable to those of human output. Nevertheless, an analysis of the qualitative data revealed two major areas where human input was still needed, specifically that GPT-4 omitted pragmatic markers and inaccurately interpreted idiomatic patient expressions. Also, certain cultural nuances present in the patient responses were interpreted literally rather than understood within the cultural context of their original presentation. The authors concluded that, although GPT-4 demonstrated quantitative accuracy in translating biomedical texts, there remains a need for human evaluation and editing prior to using these translations in clinical applications to ensure both clinical safety and adequate levels of nuance.

Chen et al. (2025) assessed the translation of critical care health education materials (English-Spanish, English-Chinese, and English-Ukrainian) using four different models: Google Translate, DeepL and a modified version of the GPT-4 Co-Pilot model, which was explicitly designed to accommodate critical care health education materials. The results were very similar to the studies mentioned above: BLEU scores showed moderate literal overlap among the four models (ranging from 33.7 BLEU for Google to 38.2 BLEU for DeepL); whereas semantic COMET scores were markedly better (ranging from 0.76 (Google) to 0.89 (GPT-4 Co-Pilot)). Moreover, the Pearson correlation coefficient between human adequacy ratings and

COMET scores was extremely high ($\rho \approx 0.91$), thereby validating the utility of neural metrics as reliable measures of the meaning of a given translation. Human raters valued GPT-4 Co-Pilot's translations nearly equivalent to those produced by a human reference (average adequacy rating = 4.6/5; average adequacy rating for human = 4.8/5). While GPT-4 Co-Pilot achieved near-human scores, it would sometimes misrender abbreviations (e.g., "BP" would be interpreted as "blood plasma" instead of "blood pressure"); similarly, it had difficulty rendering idiomatic or context-dependent phrases. Chen et al. (2025) conclude that, while advanced AI systems can perform reasonably well at translating sentences individually, they will need to be provided with additional domain-specific contextual information in order to successfully resolve ambiguous acronyms or tailor their phraseology to the particular context, both of which require a level of expertise that is currently available only through human input.

Fu and Liu (2024) compared ChatGPT-3.5-generated English-Chinese translations of scientific article abstracts with those produced by human graduate students, using various automated evaluation metrics. BLEU scores (ChatGPT = 39.8; human = 42.3) and COMET (0.86 vs. 0.89) showed minimal difference. These results indicated that the ChatGPT-generated translations were highly similar in meaning to those provided by the graduate students. Furthermore, ChatGPT's translations of technical terminology were remarkably consistent, correctly translating 96% of technical terms, which is slightly higher than humans, who achieved a 92% accuracy rate. While the ChatGPT-generated translations were generally accurate, the graduate student-generated translations tended to be more readable and adhered to the conventions of scientific writing more effectively. Specifically, many of the graduate student-generated translations were revised to improve sentence-level clarity (the graduate students revised about 2/3 of the sentences in the original abstracts because they were too long or overly complex). In contrast, most of the sentences in the ChatGPT-generated translations remained unchanged from the corresponding sentences in the source abstracts. While ChatGPT produced accurate drafts, it lacked the rhetorical awareness to enhance flow or emphasise key ideas, leading the authors to recommend a hybrid workflow: AI for draft generation, followed by human refinement.

Yan et al. (2024) conducted a large-scale study using the MQM error rating system, with professional translators assessing GPT-4's outputs against those of humans on three document types: news articles, technical documentation,

and biomedical literature. The results indicated that GPT-4 could work at a similar level to junior translators in most specialised domains and that GPT-4's technical and biomedical outputs could achieve near mid-level translator performance levels. GPT-4, however, failed to replicate the performance of senior translators working with general news; it did not capture cultural nuances or implied contextual elements to the same standard as its human counterparts. It is also worth noting that Yan et al. (2024) have identified that GPT-4's literal translation method can help prevent incorrect translations. GPT-4 correctly translated "entering his second year" into English; in contrast, some human translators incorrectly interpreted the exact phrase to imply a person who is two years old. While humans use their flexible ability for interpretation, AI systems are often biased to produce the literal translation of the source language, which can be beneficial in some cases and act as a safeguard. Overall, LLMs excel in structured, technical tasks but lag in context-rich, open-ended texts.

Table 2 summarises the findings of the studies above, illustrating how advanced AI systems now rival human quality in many areas while key differences remain.

Domain	Best AI system	Key metric (COMET/METEOR)	Human adequacy score	Distinctive findings
Legal (Briva-Iglesias et al., 2024)	GPT-4	COMET 0.82	4.5/5	High contextual accuracy despite lower BLEU than baseline MT
Biomedical (Lu et al., 2025)	GPT-4	METEOR 0.81	4.4/5	Near-human quality; minor omissions of pragmatic markers
Health Education (Chen et al., 2025)	GPT-4 CoPilot	COMET 0.89	4.6/5	Strong semantic correlation ($\rho \approx 0.91$) with human judgment
Scientific (Fu & Liu, 2024)	ChatGPT-3.5	COMET 0.86	4.3/5	Accurate terminology; weaker stylistic control

Table 2. Comparative quality of AI-generated and human translations across specialised domains.

While the quantitative gap between AI and human translation has narrowed, substantial differences remain in the qualitative dimensions, where human translators consistently outperform AI. Qualitative dimensions include pragmatic appropriateness, register and tone control, intra- and inter-sentence coherence and creative reformulation. These qualitative dimensions are challenging to quantify and are not fully captured by automatic metrics such as BLEU, METEOR, or COMET.

4.2. Human vs. AI: Qualitative differences and collaboration

Although LLM-based translators are improving in terms of accuracy and fluency, there is evidence that humans and AI each have their own unique strengths and weaknesses, suggesting a potential optimal integration of both. In general, human translators are superior at using contextual reasoning, adapting to different cultures and registers, and creatively reformulating idioms in a natural manner. Human translators can understand implied meanings, select wording suitable for an audience/context, and reorganise text for clarity or stylistic effect. On the other hand, AI systems (such as LLMs) are better at consistency, rapid production of text, and the use of precise technical terms. They consistently utilise the same technical vocabulary throughout a document and avoid simple errors (such as typos). Moreover, they can process significant amounts of text quickly and produce a first draft that typically contains no grammatical errors and utilises all technical terms correctly.

This complementarity supports a cooperative or hybrid workflow. Fischer and Läubli (2020) conducted a blind experiment comparing human and MT outputs post-edited by experts unaware of the source. The results showed that the post-editors made approximately the same number of corrections in the AI-generated translations as they did in the human-generated translations. Additionally, the total editing time was not statistically different in two out of three languages. In essence, if a well-trained domain model is used for the AI component of the translation process, the initial product can be of similar quality to that produced by a human translator, thereby making the post-editing time comparable.

Further, AI boosts productivity in translation workflows. Turner et al. (2014) studied human translation vs. a hybrid approach (combining MT with human post-editing) for translating public health documents. The hybrid workflow increased translation productivity by about 200%. At the same time, the bilingual evaluators indicated that the quality of the final translations was equivalent to that of fully human translations. These findings suggest AI augments rather than replaces human expertise. The AI system rapidly generates a reasonable first draft, and the human translator focuses on refining the technical terms, ensuring the translation is culturally accurate, and revising any minor inaccuracies in meaning or tone.

As a result, many organisations, such as the European Patent Office and the World Health Organization (WHO), have implemented new evaluation and

translation procedures that incorporate the use of both AI and human expertise. This two-stage process combines the scalability and speed of AI-based tools with the interpretive capabilities of human review. Furthermore, the use of methodological triangulation (i.e., AI metrics + human review) enhances the reliability of the translation process while reducing costs. Most importantly, the use of human review does not diminish the level of nuance that only human experts can bring to the translation process.

5. Redefining the translator's role in the AI era

While AI can produce routine translations with high fluency, human expertise remains indispensable. According to Hestness et al. (2019), modern MT systems typically reach about 80-90% of human-level quality in terms of translation. Thus, approximately 10-20% of the translated content will require a skilled human translator to review and approve the work. While AI can generate draft versions of standardised or repetitive documents (e.g., contracts, technical documentation) at a rapid pace, it often lacks a deep contextual understanding of the material being translated, which is needed to handle idiomatic expressions, cultural references and nuances of tone. As such, industry analysis suggests that even high-quality MT output generally undergoes human editing to guarantee the desired quality and nuance. Recent studies show substantial improvements in MT quality, but human post-editing is still required to reach professional standards (Orrego-Carmona, 2022; Sánchez-Torrón & Koehn, 2016).

5.1. Hybrid workflows boost productivity while keeping quality high

For many organisations, hybrid workflows offer the benefits of both the speed and efficiency of the technology as well as the accuracy and nuance that human professionals can provide. Hybrid workflows enable organisations to significantly increase the amount of work they can complete within each timeframe. Slator (2022) reports that when translators use AI drafts to produce translations and then perform post-editing, they can produce roughly twice the amount of translated material than they would without the aid of AI. A report referenced in Slator (2022) stated that after implementing MT post-editing processes into their organisation's workflow, translators were able to increase their daily word count from around 3,000 words to 7,000 words.

Industry leaders describe that "the MT draft provides a robust foundation", and then "translators review the machine-generated text to correct inaccuracies, ensuring the final translation is accurate, fluent, consistent, and adapted to context and cultural nuances" (Phrase, 2023). While larger-scale projects may add a third quality-control step, most commercial workflows rely on a two-step process: an AI draft followed by human review. This collaborative approach enables organisations to maximise the benefits of each of these technologies.

Industry-specific examples of how organisations have successfully implemented this hybrid approach show that:

- MT is used as the primary means of generating an initial draft and then refining and editing the translation to reflect the nuances and idioms of the source text.
- Productivity can increase up to 100% compared to using human-only translation processes (Slator, 2023).
- Post-edited MT output can result in translations that are equivalent in quality to those produced by human translators (Phrase, 2023).

5.2. The evolving role of human translators

The role of professional translators is changing from being the sole authors to becoming editors and supervisors. In today's AI-driven world, translators are increasingly acting as collaborative partners with AI and typically focusing on higher-order tasks (Jiménez-Crespo, 2025). The current translation process can be compared to that of a language editor or a content specialist; instead, they are tasked with reviewing and ensuring the quality and consistency of machine-generated drafts; primarily assessing their clarity, tone and overall coherence (Jiménez-Crespo, 2025).

In fact, all systems based on AI inherently have shortcomings, such as limited nuance, ethical judgment and contextual understanding (Nasir et al., 2024). This is where humans come into play to fill the missing gap; they offer the interpretation of the underlying text's meaning, resolve ambiguity and adapt idioms. For instance, although an MT system may produce a variable translation for a medical term, a human will ensure the use of the same term consistently throughout (for example, the term *reflujo ácido* will always be used when referring to *acid reflux* in Spanish). Also, as translators, they can assess

compliance with ethics, regulations and other applicable rules/guidelines that an AI system may not understand. In technical fields, they review compliance with existing standards and new terminology, while in creative or sensitive areas, they maintain the tone/intent of the original author. Human translators add contextual understanding and culturally nuanced details to the translations (Jiménez-Crespo, 2025; Vieira, 2019). Key translator functions now include:

- Terminology curator: ensuring correct and consistent use of specialised terminology; human expertise far exceeds AI's propensity to vary between synonyms (Bowker, 2020).
- Cultural mediator: adaptation of idioms, metaphors and tone to ensure the message is understood correctly in the target culture; a capability that machines do not yet possess (Cadwell et al., 2016).
- Quality gatekeeper: verification of the accuracy and fluency of the final translated text, and identification of potential minor discrepancies/stylistic flaws introduced by the AI system (Vieira, 2019; Jiménez-Crespo, 2025).

5.3. Adaptation in education and industry

With rapid changes in both the use of and the capabilities of AI MT, there has been an increasing need for translator educators to adapt their teaching methods to address these changes. Regarding the practical application of these developments, many training programmes currently offer courses on working with AI. Many of these courses provide students with the knowledge to perform the post-editing of MT, design and create appropriate prompts for LLMs, manage and use glossaries and data, and apply various types of evaluation methodologies. Industry associations have also taken notice of these changes and stress that translators need to acquire new technical skills related to AI and AI literacy. One language industry leader recently commented that "we [translators] need to train ourselves to think, speak and educate others about MT in a way that includes MT engine training, asset management, and artificial intelligence" (Jiménez-Crespo, 2025). A significant number of the new technical skills that translators will need to develop to work effectively with AI will involve teaching AI models how to perform tasks using glossaries, style guides, and structured prompts.

The emphasis in current professional guidelines is on the importance of human oversight in AI. For example, the American Translators Association (ATA) states in its current Code of Ethics that "machine translation should not be used without the ongoing involvement of professional translators" (ATA, 2018). Consistent with this view, most language industry leaders, at conferences and in professional forums, believe that AI is not intended to replace translators but to assist them. Therefore, the prevalent view is of a human-in-the-loop model where translators work in collaboration with AI systems, providing context-specific judgment to ensure quality (Slator, 2023). The major translation platforms and international organisations that provide AI-based workflows now incorporate human quality control checks into those workflows. For example, the WHO uses AI to supplement human translators by having them review MT drafts produced by AI systems, as well as draft translations created using computer-assisted translation tools and then producing final versions of the translated content (WHO, 2024).

Reports from the industry continue to indicate that a human-in-the-loop model is crucial for achieving reliable results when using AI in translation. This is evident in two recent reports from Slator (2022) and Jiménez-Crespo (2025); these practices reflect:

- New skill set is required: today's translators are expected to be proficient in technology, including the use of MT engines, translation memory management, and designing and creating prompts for AI.
- Quality standards are evolving: to improve quality, organisations are developing protocols that require the combination of automated metrics and human review. For example, the WHO's Multilingual Data Portal indicates that while MT increases coverage, the organisation cannot guarantee the accuracy of translations and therefore requires subsequent revisions (WHO, 2024).
- Human judgment is paramount: although AI can assist translators in several ways, ultimately it is still human translators who serve as the final checkpoint for quality. ATA has stated that MT can only be used with the ongoing involvement of professional translators (ATA, 2018).

5.4. Embracing a synergistic partnership

While there is concern that AI will lead to high levels of job loss, it has instead led to the evolution of the translator's job in a more positive direction. Translators are no longer responsible for translating large amounts of repetitive material. Instead, they focus on those areas where humans provide the best insight. Thus, translators are now trained professionals specialised in specific fields who edit translations to meet the desired style for the target audience and consistently enforce terminology usage throughout their assignments. Early studies, along with recent industry reports, reflect this shift. Several recent surveys suggest that the quality of machine-translated material is approaching that of professionally translated materials, and many companies are now utilising both human and AI-based, or hybrid processes, to enhance their productivity (Jiménez-Crespo, 2025; Slator, 2023). At the same time, studies indicate that although AI can produce a substantial amount of material requiring little to no additional editing, much of the final product still depends on human judgment and interpretation to achieve a professional level of quality (Orrego-Carmona, 2022; Vieira, 2019).

Current studies and evolving professional standards will continue to define the best use of collaborative efforts between human translators and AI. There appears to be evidence supporting a hybrid model being the most productive method: AI provides speed and consistency in terms of terminology usage, while human translators bring culturally relevant understanding and stylistic nuances to the table (Jiménez-Crespo, 2025; Vieira, 2019). Sadigova (2025) describes the translator as "an enhancer", someone who takes the machine-generated output and transforms it into clear and contextually relevant communication. Ultimately, the partnership formed between AI and human translators is a mutually beneficial relationship that yields increased productivity and accuracy, while also preserving the creative and nuanced aspects of language that are inherently produced through human interpretation (Jiménez-Crespo, 2025; Vieira, 2019).

6. Conclusions

There is an emerging symbiosis between machine learning and translation technologies, rather than competition or replacement for human translators

(Chen, 2024). AI enhances translators' ability to analyse and think creatively in both translation and interpretation by increasing their access to information (Chen, 2024). Nevertheless, high-level content analysis requires the application of human judgment (Schmitt, 2019). With AI reducing time on drafting translations, translators can focus on evaluating, providing creative mediation and making ethically informed decisions during the translation process (Keles et al., 2024; Yang et al., 2023). AI provides the means to develop higher-quality multilingual communication by providing better context, greater emotional connection and enhanced cultural awareness (Keles et al., 2024; Yang et al., 2023).

Looking ahead, the focus of the translation profession is expected to shift from language production to strategy design (ATA, 2018; Jiménez-Crespo, 2025). Translators will need to develop new skills to support the use and integration of AI and other technologies, such as creating validation methods, managing domain knowledge and training AI systems using human feedback (ATA, 2023; Jiménez-Crespo, 2025). The translators' role will evolve into that of a consultant, evaluator, educator and designer of the architecture of the translation process (ATA, 2023; Jiménez-Crespo, 2025). Translators will no longer be end-users of translation technologies, but rather active designers of AI-based translation systems that function ethically, reliably and inclusively (Jiménez-Crespo, 2025).

Future research should aim to develop long-term sustainable frameworks for human-in-the-loop translation processes, on creating clear guidelines and governance structures for the use of AI-based generation tools, and on rethinking the way we teach translators to include data literacy, prompt design and critical evaluation of AI-based translation systems (Chen, 2024). If developed and used appropriately and responsibly, AI will not diminish the importance of translators; it will amplify it and create the conditions in which translators will be able to act as key mediators of intercultural understanding in our increasingly technological-driven human world (Chen, 2024).

Article history:
Received 13 November 2025
Received in revised form 24 November 2025
Accepted 24 November 2025

References

Altakhaineh, A. R. M., Alghathian, G., & Jarrah, M. M. (2025). A comparative study of accuracy in human vs. AI translation of legal documents into Arabic. *International Journal of Language & Law (JLL), 14*, 63-80. https://doi.org/10.14762/jll.2025.063

American Translators Association. (2018). *ATA position paper on machine translation: A clear approach to a complex topic.* https://www.atanet.org/advocacy-outreach/ata-position-paper-machine-translation-a-clear-approach-to-a-complex-topic

American Translators Association. (2023). *MTPE - Overview of prevailing industry guidelines.* https://www.ata-divisions.org/LTD/mtpe-overview-of-prevailing-industry-guidelines

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International conference on learning representations (ICLR)*. *arXiv.* https://arxiv.org/abs/1409.0473

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65-72). Association for Computational Linguistics.

Bowker, L. (2020). *Machine translation and global research: Towards improved machine translation literacy in the scholarly community.* Emerald.

Briva-Iglesias, V., Camargo, J. L. C., & Doğru, G. (2024). Large language models "ad referendum": How good are they at machine translation in the legal domain? *arXiv*. https://doi.org/10.48550/arXiv.2402.07681

Cadwell, P., Castilho, S., O'Brien, S., & Mitchell, L. (2016). Human factors in machine translation and post-editing among institutional translators. *Translation Spaces, 5*(2), 222-243. https://doi.org/10.1075/ts.5.2.04cad

Chen, C. L., Dong, Y., Castillo-Zambrano, C., Bencheqroun, H., Barwise, A., Hoffman, A., Nalaie, K., Qiu, Y., Boulekbache, O., & Niven, A. S. (2025). A systematic multimodal assessment of AI machine translation tools for enhancing access to critical care education internationally. *BMC Medical Education, 25*(1), 1022. https://doi.org/10.1186/s12909-025-07452-9

Chen, P. (2024). The impact of generative AI on the role of translators and its implications for translation education. *Education Insights*, *1*(2), 24-33. https://doi.org/10.70088/kc9vk395

Fischer, L., & Läubli, S. (2020). What's the difference between professional human and machine translation? A blind multi-language study on domain-specific MT. *arXiv.* https://arxiv.org/abs/2006.04781

Freitag, M., Rei, R., Mathur, N., Lo, C.-K., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., & Martins, A. F. T. (2022). Results of WMT22 metrics shared task: Stop using BLEU—Neural metrics are better and more robust. In *Proceedings of the seventh conference on machine translation (WMT)* (pp. 46-68). Association for Computational Linguistics.

Fu, L., & Liu, L. (2024). What are the differences? A comparative study of generative artificial intelligence translation and human translation of scientific texts. *Humanities and Social Sciences Communications, 11*(1), 1-12. https://doi.org/10.1057/s41599-024-03726-7

Hassan, H., et al. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv.* https://arxiv.org/abs/1803.05567

Hestness, J., Ardalani, N., & Diamos, G. (2019). Beyond human-level accuracy: Computational challenges in deep learning. In *Proceedings of the 24th symposium on principles and practice of parallel programming* (pp. 1-14). Association for Computing Machinery. https://doi.org/10.1145/3293883.3295710

Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation.* Academic Press.

Jiménez-Crespo, M. A. (2025). Human-centered AI and the future of translation technologies: What professionals think about control and autonomy in the AI era. *Information, 16*(5), 387. https://doi.org/10.3390/info16050387

Keles, B., Gunay, M., & Caglar, S. I. (2024). LLMs-in-the-loop Part 1: Expert small AI models for biomedical text translation. *arXiv.* https://arxiv.org/abs/2407.12126

Kocmi, T., & Federmann, C. (2023). GEMBA-MQM: Detecting translation quality error spans with GPT-4. *arXiv.* https://arXiv:2310.13988

Koehn, P. (2010). *Statistical machine translation.* Cambridge University Press.

Lommel, A. R., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica, 12*, 455-463.

https://doi.org/10.5565/rev/tradumatica.77

Lu, S. C., Xu, C., Kaur, M., Edelen, M. O., Pusic, A., & Gibbons, C. (2025). Can machine translation match human expertise? Quantifying the performance of large language models in the translation of patient-reported outcome measures (PROMs). *Journal of Patient-Reported Outcomes, 9*(1), 94. https://doi.org/10.1186/s41687-025-00926-w

Mohamed, Y. A., Khanan, A., Bashir, M., Mohamed, A. H. H., Adiel, M. A., & Elsadig, M. A. (2024). The impact of artificial intelligence on language translation: A review. *IEEE Access, 12*, 25553-25579. https://doi.org/10.1109/ACCESS.2024.3366802

Moneus, A. M., & Sahari, Y. (2024). Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon*, *10*(6), e28106. https://doi.org/10.1016/j.heliyon.2024.e28106

Nasir, S., Khan, R. A., & Bai, S. (2024). Ethical framework for harnessing the power of AI in healthcare and beyond. *IEEE Access, 12*, 31014-31035. https://doi.org/10.1109/ACCESS.2024.3369912

Nazi, Z. A., & Peng, W. (2024). Large language models in healthcare and medical domain: A review. *Informatics, 11*(3), 57. https://doi.org/10.3390/informatics11030057

OpenAI. (2023). *GPT-4 technical report*. *arXiv.* http://arxiv.org/abs/2303.08774

Orrego-Carmona, D. (2022). Machine translation in everyone's hands—Adoption and changes among general users of MT. *Tradumàtica, 20*, 322-339. https://doi.org/10.5565/rev/tradumatica.324

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Phrase. (2023). *Machine translation post-editing: perspectives and best practices*. https://phrase.com/blog/posts/machine-translation-post-editing-best-practices

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation (WMT15)* (pp. 392-395). Association for Computational Linguistics.

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.

Sadigova, N. (2025). The role of artificial intelligence in modern-term formation. *Path of Science, 11*(3), 7006-7012. https://doi.org/10.22178/pos.115-26

Sánchez-Torrón, M., & Koehn, P. (2016). Machine translation quality and post-editor productivity. In *Proceedings of the 12th conference of the association for machine translation in the Americas (AMTA 2016), vol. 1: MT researchers' track* (pp. 16-26). Association for Machine Translation in the Americas https://aclanthology.org/2016.amta-researchers.2.pdf

Schmitt, P. A. (2019). Translation 4.0—Evolution, revolution, innovation or disruption? *Lebende Sprachen, 64*(2), 193-229. https://doi.org/10.1515/les-2019-0013

Slator. (2023). *How fast can you post-edit machine translation?* https://slator.com/how-fast-can-you-post-edit-machine-translation

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th conference of the association for machine translation in the Americas (AMTA 2006)* (pp. 223-231). Association for Machine Translation in the Americas.

Turner, A. M., Bergman, M., Brownstein, M., Cole, K., & Kirchhoff, K. (2014). A comparison of human and machine translation of health promotion materials for public health practice: Time, costs, and quality. *Journal of Public Health Management and Practice, 20*(5), 523-529. https://doi.org/10.1097/PHH.0b013e3182a95c87

Vashee, K. (2019). Understanding machine translation quality: BLEU scores. *RWS Blogs.* https://www.rws.com/blog/understanding-mt-quality-bleu-scores

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in neural information processing systems 30 (NIPS 2017)*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vieira, L. N. (2019). Post-editing of machine translation. In M. O'Hagan (Ed.), *The Routledge handbook of translation and technology* (pp. 319-335). Routledge.

World Health Organization. (2024). *Language*

translations. https://data.who.int/about/datadot/translations

Yan, J., Yan, P., Chen, Y., Li, J., Zhu, X., & Zhang, Y. (2024). GPT-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv*. https://doi.org/10.48550/arXiv.2407.03658

Yang, X., Zhan, R., Wong, D. F., Wu, J., & Chao, L. S. (2023). Human-in-the-loop machine translation with large language model. In *Proceedings of the 24th machine translation summit (MT Summit 2023), vol. 2: Users track* (pp. 120-133). Asia-Pacific Association for Machine Translation. https://aclanthology.org/2023.mtsummit-users.8

Zaim, M., Arsyad, S., Waluyo, B., Ardi, H., Al Hafizh, M., Zakiyah, M., Syafitri, W., Nusi, A., & Hardiah, M. (2025). Generative AI as a cognitive Co-Pilot in English language learning in higher education. *Education Sciences*, *15*(6), 686. https://doi.org/10.3390/educsci15060686

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1904.09675

Zhong, T., et al. (2024). Evaluation of OpenAI o1: Opportunities and challenges of AGI. *arXiv*. https://doi.org/10.48550/arXiv.2409.18486

Pascual Cantos-Gómez is a Full Professor of English Linguistics at the University of Murcia, where he has been a faculty member since 1990. He is an internationally recognised specialist in corpus linguistics and empirical, data-driven approaches to language analysis. As director of the LACELL Research Group, he has led advanced research in computational lexical semantics, corpus-based modelling of collocational structure, stochastic and multivariate methods for semantic disambiguation, and quantitative analyses of linguistic patterns associated with cognitive and clinical conditions. His work also includes statistical modelling of deceptive discourse and the development of intelligent systems for managing lexical-semantic and ontological knowledge. He served as President of the Spanish Association of Corpus Linguistics (AELINCO) from 2014 to 2023. His scholarly production exceeds 150 publications, including monographs, edited volumes and articles with major academic publishers such as Routledge, Benjamins, De Gruyter, Springer, Bloomsbury and Georgetown University Press. Recent corpus-oriented contributions include co-editing *The Routledge Handbook of Spanish Corpus Linguistics* (2022). He has delivered over 200 invited lectures, seminars and conference presentations.