Innovative Research Journal of **Artificial Intelligence** (IRJAI)
DOI: https://doi.org/10.62497/irjai.275

Review Article

Artificial Intelligence and Machine Translation: ChatGPT, Google Translate, DeepL, Human-AI Collaboration, Post-Editing, and Translation Quality

Esraa Abd Elnaser Mohamed Hassan[1]* and Bushra Naseer[2]
[1] *College of Foreign Languages and Cultures, Chengdu University of Technology, 610059 Chengdu, China*
[2] *Department of English Linguistics, University of Wah, Wah, Pakistan*

Citation: Hassan EAEM, Naseer B. Artificial Intelligence and Machine Translation: ChatGPT, Google Translate, DeepL, Human-AI Collaboration, Post-Editing, and Translation Quality. Innov. Res. J. AI. 2026;4(1):1-14. Available from: https://irjpl.org/irjai/article/view/275

Article Info
Received: Oct 10, 2025
Revised: Jan 19, 2026
Accepted: Feb 02, 2026

Keywords
artificial intelligence, machine translation, ChatGPT, Google Translate, DeepL, large language models, post-editing, human–AI collaboration, translation quality

Copyright © 2026 The Author(s).

Published by Innovative Research Journals (IRJPL).

This is an Open Access article under the CC BY NC 4.0 license. This license enables reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.

Abstract

Introduction: Artificial intelligence has transformed machine translation from rule-based and statistical systems to neural and large language model systems that can produce fluent multilingual text at unprecedented speed. ChatGPT, Google Translate, and DeepL represent different manifestations of this transformation, while human–AI workflows increasingly combine automated generation with expert revision.

Objective: This paper examines the development of AI-based machine translation and compares general-purpose large language models, translation-oriented platforms, human translation, and hybrid workflows. It focuses on post-editing, quality evaluation, hallucination, contextual accuracy, terminology, and the conditions under which AI output can be considered fit for purpose.

Method: An integrative literature review was conducted using foundational studies on neural machine translation and Transformer architectures, empirical studies of multilingual large language models, research on post-editing productivity, automatic and human evaluation, hallucination, and professional standards. Contemporary research from 2024–2025 was emphasized where available.

Results: The synthesis indicates that AI translation provides major gains in speed, scalability, consistency, and drafting capacity, but quality varies by language pair, domain, text type, and system. Large language models provide unusually strong instruction-following and interactive revision, while specialized translation systems remain valuable for high-volume and repeatable tasks. Human review is essential when semantic precision, terminology, cultural appropriateness, or accountability is important.

Conclusion: AI is best understood as a translation partner rather than a universal substitute for professional translators. The most reliable workflow uses AI for candidate generation, comparison, terminology support, and repetitive work while assigning humans responsibility for interpretation, error detection, cultural judgment, and final approval.

** Corresponding Author:*
Esraa Abd Elnaser Mohamed Hassan
Email: hs5932457@gmail.com

1. Introduction

Machine translation (MT) has moved from an experimental computational-linguistics problem to an everyday communication technology. People routinely use automated translation to read websites, communicate across languages, understand documents, and produce multilingual content. The technological trajectory behind this transformation has been long and cumulative. Early rule-based systems encoded grammatical and lexical knowledge explicitly; statistical machine translation learned translation correspondences from parallel corpora; neural machine translation then replaced many manually engineered components with learned representations. The result has been a shift from phrase-level substitution toward systems capable of modeling broader relationships within sentences and documents. The modern period is especially significant because translation is now being incorporated into general-purpose artificial intelligence. Instead of a user selecting a language pair and receiving a single output, a conversational model can be instructed to preserve tone, simplify terminology, adapt register, produce alternatives, explain ambiguity, or revise a previous translation. This changes the practical definition of machine translation. Translation becomes one operation within a broader language workflow involving summarization, rewriting, terminology management, explanation, and quality control. Neural machine translation established much of the technical foundation for this change. Attention-based encoder-decoder models allowed systems to learn which source tokens were most relevant to each target token, improving the treatment of long-distance relationships [1]. Google's Neural Machine Translation system showed that neural approaches could be deployed at industrial scale and substantially improve translation performance over earlier phrase-based approaches [2]. The Transformer architecture then replaced recurrence with self-attention as the central mechanism, enabling greater parallelization and becoming a foundation for later language models [3].

The rise of large language models (LLMs) extends this development. GPT-4 demonstrated strong general language capabilities and broad multilingual performance while also documenting important limitations in reliability and factual behavior [4]. Research specifically examining multilingual machine translation has shown that LLMs can be competitive with strong supervised systems in some directions but can still lag behind commercial translation systems, particularly for low-resource languages [5]. This variation is central to any responsible assessment of AI translation: high fluency in one language pair cannot be generalized to all language pairs. Contemporary translation practice therefore involves several different technological categories. ChatGPT is primarily a conversational LLM that can perform translation as an instructed task. Google Translate is a large-scale translation service with a long history of specialized MT research and deployment. DeepL is a translation-oriented commercial platform that emphasizes high-quality output and professional workflows. Human translators, by contrast, perform interpretive work that combines language knowledge with domain expertise, cultural awareness, stylistic judgment, and accountability.

The practical question is consequently not whether AI can translate. It can. The more important question is what kind of translation a system can reliably produce, how its output should be evaluated, how much human effort is required to correct it, and which tasks should remain under direct human control. This paper addresses these questions through a synthesis of technical, empirical, and professional literature.

Objective

The primary objective of this review is to develop a balanced academic account of contemporary AI-assisted translation. The discussion deliberately avoids a simple competition between humans and machines because the systems perform different functions and are optimized for different environments. The specific objectives are to: (1) explain the technological development from neural machine translation to LLM-based translation; (2) compare ChatGPT, Google Translate, DeepL, human translation, and hybrid workflows; (3) examine the major dimensions of translation quality; (4) analyze post-editing effort and productivity; (5) investigate hallucination and other reliability problems; (6) assess the value and limitations of automatic metrics; and (7) propose a practical human–AI workflow for

professional translation.

2. Literature Review

Statistical machine translation treated translation as a probabilistic inference problem over learned correspondences. Its strength was the ability to learn from large parallel corpora without manually encoding every translation rule. However, statistical systems often struggled with long-range dependencies, word-order differences, and coherent discourse. Neural translation changed the representational model by learning continuous vector representations and end-to-end mappings between source and target sequences. Attention mechanisms were especially influential because they allowed the decoder to focus on different parts of the source sentence as it generated each target token [1]. Instead of compressing the entire source into a single fixed representation, the model could dynamically access relevant information. This improved translation of longer and more structurally complex sentences.

GNMT demonstrated how neural architectures could be adapted to a production translation service [2]. Its importance was not only the reported quality improvement but also the demonstration that neural translation could function as infrastructure for millions of users. This established a path toward the current expectation that translation should be instantaneous and available across large numbers of language pairs. The Transformer subsequently introduced a more efficient attention-based architecture [3]. Its ability to process sequences in parallel contributed to the rapid scaling of language models. Many contemporary LLMs inherit the Transformer family of ideas, although commercial systems differ in architecture, training data, post-training methods, and deployment details.

LLMs introduce a broader task representation. The model does not need to interpret an instruction as merely a request for word substitution. It can treat translation as a constrained generation task. This creates opportunities for register control, terminology instructions, audience adaptation, and interactive revision. At the same time, the generative nature of LLMs introduces risks: a system may produce a plausible translation that changes, omits, or invents information. The distinction between fluency and fidelity therefore becomes increasingly important.

ChatGPT and LLM-Based Translation

ChatGPT-class systems have changed translation practice because the interface is conversational. A user can ask for a translation, inspect the result, request a more formal version, ask why a phrase was interpreted in a particular way, and then request a revised version. This iterative process can be especially useful for writers and translators who are exploring multiple target-language formulations. Instruction following is a major advantage. A prompt can specify the target language, regional variety, audience, tone, terminology, formatting, and translation philosophy. Research on multilingual LLMs shows that translation ability is influenced by language resources and linguistic relationships, while instruction-based multilingual fine-tuning can substantially improve translation behavior even for some language pairs not directly seen during instruction tuning [6]. However, instruction following should not be confused with guaranteed obedience to the source text. A model may interpret an ambiguous phrase, select a plausible sense, and generate a coherent sentence without signaling uncertainty. In a translation context, this can be dangerous because the output may look more trustworthy precisely because it is fluent.

Research comparing LLMs across many language directions has found strong potential but also substantial gaps between LLMs and commercial translation systems, especially in low-resource settings [5]. This indicates that general language capability does not automatically imply uniform translation quality. Performance must be tested at the language-pair and domain level. ChatGPT is therefore particularly useful for translation tasks in which interaction itself creates value: producing several stylistic candidates, explaining ambiguous phrases, adapting register, transforming drafts, and helping a translator reason through difficult alternatives. It is less appropriate to treat an unreviewed response as authoritative in high-stakes or publication-critical translation.

Google Translate and DeepL as Specialized Translation Platforms

Google Translate and DeepL represent translation-oriented services rather than purely general-purpose conversational models. Their importance lies partly in the engineering of translation as a production service: language-pair coverage, latency, document processing, terminology, user interfaces, and large-scale deployment all matter alongside raw model quality. Google's research history demonstrates the transition from phrase-based to neural translation [2]. More recent work shows that the boundary between specialized MT and LLMs is becoming less distinct. In its WMT2025 research submission, Google described a system built from a large language model that was supervised-fine-tuned with high-quality synthetic parallel data and then optimized using reinforcement learning and translation-oriented reward models [15]. The study also reported a controllable trade-off between more literal and more creative translation behavior. This illustrates a broader industry trend: specialized translation services are increasingly adopting LLM techniques rather than remaining architecturally separate from them. DeepL likewise emphasizes translation quality and human comparative evaluation. Its current quality methodology describes blind comparative human judgment and discusses limitations of automatic metrics such as BLEU and COMET [16]. Such information is valuable for understanding how a commercial provider thinks about quality, but it should be interpreted carefully because vendor-produced benchmarks are not equivalent to independent multi-system evaluations.

For practical users, the distinction between Google Translate and DeepL is less important than the principle that a translation service should be selected according to the task. A high-volume customer-support workflow may prioritize speed and coverage. A technical localization project may prioritize terminology and consistency. A literary text may require a system that preserves unusual syntax and imagery. A single global ranking cannot capture all of these needs.

Comparative Characteristics of Major Translation Approaches

Before comparing systems, it is useful to distinguish system specialization from general language capability. ChatGPT can translate but is designed for a wide range of language tasks. Google Translate and DeepL are strongly oriented toward translation services. Human translators are not software systems at all; they combine linguistic interpretation with professional judgment. Hybrid workflows combine these capabilities rather than selecting only one.

The comparison in Table 1 is therefore functional rather than a universal quality ranking. It focuses on the type of value each approach can provide in a real translation workflow.

Table 1: Functional comparison of major translation approaches

Approach	Primary orientation	Major strengths	Main risks or limitations	Best-fit applications
ChatGPT / LLM	General language generation and instruction following	Context adaptation; stylistic control; explanations; alternative formulations; interactive revision	Prompt sensitivity; hallucination; variable terminology; uneven multilingual performance	Drafting, multilingual communication, style-sensitive translation, exploratory work
Google Translate	Large-scale translation service	Broad accessibility; speed; scalable deployment; mature translation infrastructure	Quality varies by language and domain; contextual and critical errors can remain	Everyday translation, high-volume general content, rapid comprehension
DeepL	Translation-oriented AI service	Strong fluency; professional workflows;	Proprietary system; performance varies by language/domain; vendor	Business and technical translation, localization, document

		terminology and document features	benchmarks require independent interpretation	translation
Human translator	Human linguistic and cultural interpretation	Context; cultural knowledge; accountability; creativity; ambiguity resolution	Higher time and cost requirements	Literary, legal, medical, diplomatic, publication-critical work
Human–AI hybrid	AI generation plus human judgment	Combines speed, scalability, and expert quality control	Requires workflow design and skilled review	Professional localization, technical content, publishing, controlled enterprise workflows

Empirical findings: language-pair and domain effects

Large-scale multilingual evidence cautions against a universal ranking. Zhu et al. evaluated eight large language models across many translation directions and reported that GPT-4 outperformed the supervised NLLB baseline in 40.91% of directions, yet still lagged behind Google Translate, particularly for low-resource languages [5]. Results from high-resource pairs such as English–Spanish therefore should not be generalized automatically to language pairs with less training data.

Direct comparisons of ChatGPT, Google Translate, and DeepL also depend on text type. Gao et al. compared ChatGPT (GPT-3.5), Google Translate, and DeepL on 21 Chinese classical poems, with nine highly qualified bilingual raters scoring fidelity, fluency, and style. ChatGPT received a mean overall score of 3.61/5, compared with 2.29 for DeepL and 2.13 for Google Translate; fidelity scores were 3.83, 2.97, and 2.75, respectively [19]. A prompt asking ChatGPT to interpret a poem before translating improved style and overall ratings, but did not significantly improve fidelity. These findings support ChatGPT's promise for this specific literary task, not a general claim of superiority across domains.

A 2025 medical-text study compared DeepL, ChatGPT, and Google Translate on 500 German sentences translated into English, French, and Polish using BLEU, METEOR, and COMET. BLEU medians were 0.32–0.36 for German–English, 0.28–0.31 for German–French, and 0.23–0.26 for German–Polish; English translations generally scored higher, while most system comparisons did not show a consistent statistically significant winner across all metrics [20]. Separately, the WMT25 terminology task found that providing correct glossaries consistently improved terminology accuracy and overall quality, while document-level finance translation remained difficult [23]. Together, these findings support evaluation by language pair, domain, and document length rather than by tool branding alone.

Direct comparisons with professional translation are also mixed. In a study of 20 standardized pediatric discharge instructions translated into Spanish, Brazilian Portuguese, and Haitian Creole, ChatGPT-4 and Google Translate received ratings comparable to professional translations for Spanish and Portuguese. For Haitian Creole, professional translations scored significantly better; potentially clinically significant errors were reported in 33.3% of ChatGPT outputs and 23.3% of Google Translate outputs, compared with 8.3% of professional translations [21]. This illustrates why results from one language cannot establish equivalence to professional translation in another.

Kong et al. studied 50 sets (316 sentences) of patient-specific emergency-department discharge instructions translated from English into Spanish, Chinese, and Russian. Sentence-level accuracy for ChatGPT-4 and Google Translate was 97%/96% in Spanish, 95%/90% in Chinese, and 89%/80% in Russian. Yet at least one inaccuracy occurred in 16%, 24%, and 56% of ChatGPT instruction sets and 24%, 56%, and 66% of Google Translate sets, respectively; potentially harmful errors affected up to 6% of instruction sets [22]. This demonstrates that strong

sentence-level averages can conceal cumulative risks at document level.

Taken together, these studies do not establish a universal ordering of ChatGPT, Google Translate, DeepL, and human translation. Human translators contribute contextual judgment and professional accountability, while comparative outcomes depend on the text, language direction, evaluation method, prompt, and quality threshold. Table 1 should therefore be read as a functional comparison; organizations should choose systems through task-specific testing and risk classification rather than a single global ranking.

Translation Quality: What Does 'Good' Mean?

Translation quality is multidimensional. A target sentence can be grammatically correct but semantically inaccurate. It can preserve meaning but sound unnatural. It can be fluent and accurate but use terminology that is unacceptable in a specialist field. It can be linguistically excellent but culturally inappropriate for the intended audience. BLEU was one of the most influential early automatic metrics. It measures overlap between machine output and reference translations and made large-scale comparison relatively straightforward [7]. However, lexical overlap is not the same as meaning. Two equally valid translations may use very different words, while a translation that shares many words with a reference can still contain a critical error.

COMET represents a later generation of evaluation methods. It uses multilingual representations and information from source and target texts to predict human judgments [8]. Such learned metrics often correlate better with human evaluation than purely lexical metrics, but they still cannot completely represent literary, pragmatic, or domain-specific requirements. Human evaluation remains essential because professional translators can inspect the source and target together and identify errors that automated metrics may not capture. Freitag and colleagues demonstrated that professional evaluation with full document context can produce system rankings that differ from crowd-based evaluation [9]. Their findings emphasize that the evaluation method itself can influence the conclusion.

Translation quality should therefore be defined in relation to purpose. For a weather alert, factual accuracy and clarity dominate. For a marketing slogan, persuasive effect may matter more. For a legal contract, terminology and legal meaning are critical. For a novel, voice and literary effect may be more important than lexical similarity.

Dimensions of Translation Quality

A practical quality framework should distinguish between errors that change meaning and errors that merely reduce stylistic smoothness. Critical errors can include negation mistakes, incorrect numbers, wrong entities, omitted instructions, and invented content. Less severe issues may include awkward collocation or a slightly unnatural phrase. Terminology is particularly important in specialized translation. A medical term, engineering component, software feature, or legal concept may have a preferred target-language equivalent. An AI system can be highly fluent while choosing an unacceptable term. Terminology control should therefore be treated as a separate quality layer rather than assumed to emerge automatically from general language competence.

Document context is another major factor. A sentence can be ambiguous in isolation but clear in the full document. Pronouns, repeated terminology, character names, tense, and discourse relations often require information from earlier sentences. Recent WMT evaluation has increasingly moved toward document-level and more difficult test material, reflecting the limitations of isolated-sentence evaluation [15]. Cultural appropriateness is also important. Translators must recognize references, politeness conventions, idioms, humor, and assumptions that may not transfer directly. A machine translation that is syntactically correct can still fail pragmatically.

Finally, factual fidelity deserves explicit attention. Translation systems should not add information merely because the target-language sentence sounds more natural. The source text remains the controlling evidence. When the source is ambiguous, a system should preserve that ambiguity or signal the uncertainty rather than silently inventing a resolution.

Table 2: Translation-quality dimensions and practical evaluation methods

Dimension	Core question	Useful evaluation method	Why it matters
Adequacy	Is the intended meaning preserved?	Source-target review; COMET; MQM	Prevents semantic distortion
Fluency	Does the target text read naturally?	Human review; language-quality assessment	Ensures readability
Terminology	Are specialist terms correct and consistent?	Glossary checks; termbase review	Critical in technical domains
Context	Does the segment fit the whole document?	Document-level review	Resolves pronouns, references and discourse relations
Style and register	Is tone appropriate for the audience?	Human comparative judgment	Important in business, marketing and literary work
Cultural appropriateness	Does the target text function in the target culture?	Expert cultural review	Avoids pragmatic and cultural failures
Factual fidelity	Has anything been added, omitted or invented?	Source-target verification	Controls hallucination and critical errors
Post-editing effort	How much human work is required?	Time, edits, HTER	Connects system quality to workflow cost

Post-Editing: From Machine Output to Professional Translation

Post-editing is the process of revising machine-generated output so that it meets a defined quality standard. ISO 18587:2017 specifies requirements for full human post-editing and identifies competencies expected of post-editors [14]. The standard reflects an important professional principle: machine output is a starting point, not automatically a finished translation. Post-editing differs from ordinary proofreading because the translator must actively compare source and target. The post-editor must look for omissions, additions, mistranslations, terminology errors, grammatical problems, register shifts, and contextual inconsistencies. A polished sentence is not acceptable if it changes the source meaning. Research on HTER introduced a practical way to estimate post-editing effort by measuring the edits needed to transform machine output into a human reference [10]. This connects system quality with human workload. The concept is important because a machine system that produces attractive drafts but requires extensive correction may not deliver meaningful productivity gains.

Production studies have found that post-editing can be faster than translating from scratch in some conditions, but the benefit varies by language pair and text type [11]. Research with professional English-Spanish translators likewise found a relationship between higher MT quality and reduced post-editing time [12]. These results support a conditional rather than absolute claim: better MT can reduce effort, but poor MT can create additional cognitive work. Post-editing is most effective when the workflow identifies segments that are likely to be useful before a translator invests time in them. Quality estimation, terminology checking, and risk classification can help direct human attention toward difficult segments. In high-risk domains, the workflow should default to human verification rather than assume that a high average score guarantees safety.

Quantitative findings show why productivity claims must remain conditional. In a production study involving 20 experienced media professionals working with English–German and English–Dutch, post-editing was faster than manual translation in two of four directions—Dutch→English and, marginally, English→German—even though participants initially expected post-editing to take longer [11]. In a separate study of professional English–Spanish translators, each one-point increase

in BLEU was associated with a 0.16-second-per-word reduction in post-editing time (about 3–4%); the lowest-scoring MT system required the greatest editing effort and yielded the lowest post-edited quality [12]. These estimates depend on the systems, language directions, texts, and workflow tested and should not be treated as universal productivity rates.

Human-AI Collaboration

Human–AI translation should be understood as a division of labor. AI is particularly strong at generating candidate text quickly, comparing alternatives, identifying repeated terminology, and performing large amounts of routine language transformation. Humans are stronger at interpreting communicative purpose, resolving culturally loaded ambiguity, evaluating whether a choice is acceptable for a specific audience, and taking responsibility for the final product. The strongest workflow is therefore not 'machine first, human later' in every situation. Sometimes the human should define the terminology and translation policy before generation. Sometimes AI can propose several options for a difficult sentence, after which the human chooses. Sometimes the human may translate a key paragraph manually and use AI only for secondary material. The correct arrangement depends on risk and text type. An effective collaboration model also makes uncertainty visible. Instead of asking the AI for a single supposedly correct answer, the translator can request alternative translations and identify the reasons for the differences. This turns the model into a candidate-generation and reasoning aid rather than an authority.

Human oversight is particularly important when the source text is ambiguous, culturally specific, or legally sensitive. A system should not be allowed to transform an uncertain source into an apparently certain target without review. The human translator acts as the final semantic and pragmatic gatekeeper.

Hallucination, Omission, and Factual Risk

Hallucination is one of the most important reliability problems in generative translation. A hallucinated translation contains material that is not supported by the source, or it distorts the source so substantially that the target no longer represents the original. This can happen even when the target language is highly fluent. Research across more than 100 language pairs has shown that multilingual NMT and LLM systems can produce hallucinations, with particular difficulties in low-resource directions and in certain directions away from English [13]. The study also found qualitative differences between hallucinations produced by NMT and LLM systems. This suggests that simply increasing model size is not a complete solution. Hallucination can appear as addition, omission, substitution, or fabrication. For example, a model might insert a plausible adjective, replace an unfamiliar organization with a familiar one, or invent a relationship between two sentences. In literary contexts, hallucination may be especially difficult to detect because creative language is already expected. In technical contexts, the consequences may be more obvious and potentially serious.

Risk management therefore requires explicit source-target comparison. Important entities, numbers, dates, names, units, negations, and instructions should receive targeted verification. In high-stakes workflows, the system should also be evaluated on the specific domain rather than relying only on general benchmark results.

Terminology, Consistency, and Document Context

Terminology management is a practical area in which human–AI collaboration can be highly effective. A translator can establish a glossary before generation, provide preferred equivalents, and then ask the AI to preserve those choices. The same process can be used to identify inconsistent translations of names, product features, or technical terms. Consistency extends beyond individual terms. A long document may require stable translation of headings, pronouns, repeated concepts, and stylistic conventions. An LLM can assist by reviewing repeated passages and identifying variations, while a translation management system or glossary can provide a more deterministic source of terminology. Document-level context is particularly important for pronouns and discourse relations. A sentence that contains 'they', 'this', or 'the former' may be impossible to translate accurately without knowing what was mentioned previously. Current evaluation research increasingly tests systems on complete documents for this reason [15]. However, giving a model more context does not

guarantee that it will use the context correctly. The system can still prioritize a locally plausible interpretation. Human review is therefore needed when an ambiguity has consequences for meaning.

Table 3: Risk-based selection of human and AI involvement

Risk level	Typical content	AI role	Human role	Recommended final control
Low	Informal messages, personal notes, general web content	Primary generator	Light review when needed	Spot check
Moderate	Business correspondence, internal reports, routine localization	Drafting and terminology assistance	Systematic post-editing	Second-pass QA
High	Legal, medical, financial, safety, regulatory content	Candidate generation only where appropriate	Expert translation and verification	Independent expert review
Publication-critical	Literature, public statements, branded campaigns, academic manuscripts	Brainstorming and alternative generation	Full editorial responsibility	Final human authorization

Operational criteria for assigning risk levels

Assign the category before selecting a translation tool. Record the likely consequence of an error, whether readers will act on the text, whether errors can be reversed, the sensitivity of the terminology or data, and whether the text is public-facing or authoritative. If a task meets more than one description, use the highest applicable category. Escalate at least one level when the source is ambiguous, contains critical numbers, dates, units, negation or dosage instructions, or when the selected system has not been validated for the relevant language pair and domain.

Low risk: An error would cause only minor inconvenience and be easy to reverse; the text is for personal or internal understanding (e.g., an informal message or a non-critical webpage). AI may generate the translation, followed by spot checks for names, numbers, and omissions.

Moderate risk: An error could cause rework, delay, or routine customer confusion but is not expected to affect health, safety, legal rights, or material finances (e.g., routine business correspondence, internal reports, or non-regulated product FAQs). Use AI as a draft, then have a competent bilingual reviewer post-edit the full text and perform a second-pass quality check.

High risk: An error could affect health, safety, legal rights, finances, regulatory compliance, or operational decisions (e.g., medication directions, contract clauses, financial disclosures, or safety procedures). AI output should be supporting material only; a qualified domain translator must check the complete source and target text, followed by independent expert review before release.

Publication-critical: The translation will be published or issued as an authoritative public text, or its voice and nuance materially affect reputation, meaning, or authorship (e.g., academic manuscripts, public statements, literary works, or brand campaigns). Require professional translation/editorial ownership, source-fidelity and terminology checks, fact/reference verification, and final author/editor approval. AI may generate alternatives but should not provide final sign-off.

For implementation, the project owner should record the risk tier, language pair, selected tool and version, glossary, reviewer, acceptance criteria, and release authorization in the job brief. A high average quality score must never override an unresolved critical error.

The workflow shown in Figure 1 places quality control throughout the process rather than adding proofreading only at the end. The first step defines purpose and audience because acceptable translation quality is task-dependent. AI generation follows only after those constraints are established. Automatic checks then identify predictable issues,

while human review addresses semantic and contextual problems. Final quality assurance provides an independent verification layer.

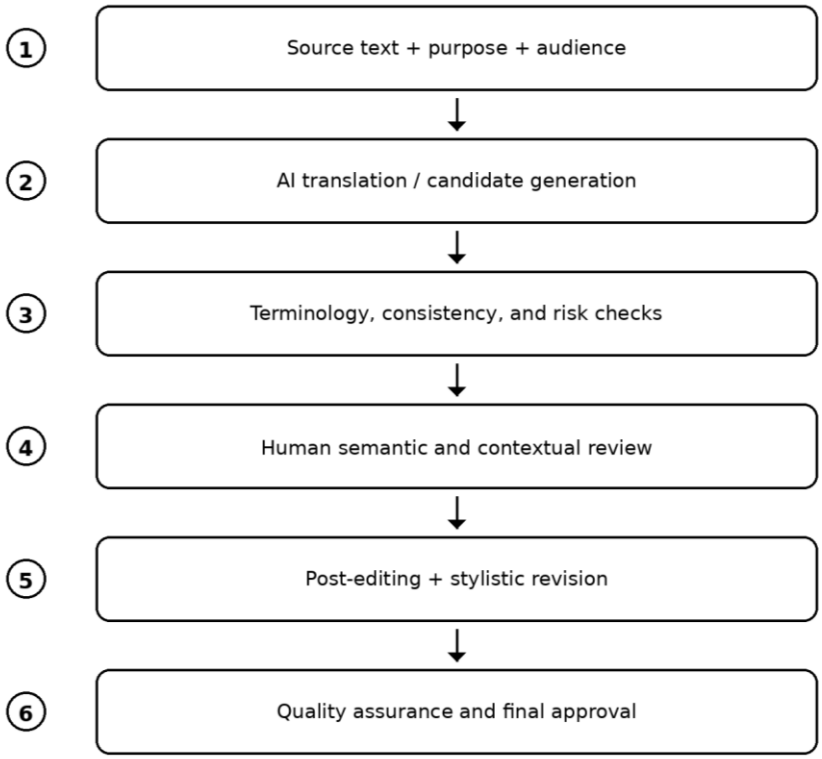


Figure 1: Human–AI translation workflow.

Figure 2 illustrates why translation quality should be evaluated in layers. The first layers address source-target fidelity and fluency. Later layers consider terminology, document coherence, culture, and error severity. The final human judgment integrates these dimensions with the purpose of the translation. This approach prevents a single automatic score from becoming the sole decision criterion.



Figure 2: Multi-layer translation-quality assurance framework.

Organizations can operationalize the figures by setting a pass/fail gate at each stage: (1) approve purpose, audience, source language, target locale, risk tier, and glossary before generation; (2) check for missing or added content and verify names, dates, numbers, units, and negation; (3) have a reviewer classify errors by severity and log post-editing time and changes; and (4) require zero unresolved critical errors and documented sign-off before release. BLEU, COMET, or a vendor score can support triage, but should not be the sole release criterion, particularly for medical, legal, or publication-critical content [9,14,20,22,23].

3. Methodology

This paper uses an integrative literature-review methodology. Foundational publications were selected to establish the development of attention-based NMT and Transformer architectures [1-3]. Studies of LLM translation were then used to examine contemporary multilingual capabilities and instruction-following behavior [4-6]. Research on BLEU, COMET, human evaluation, and error analysis was included to address the problem of measuring translation quality [7-9]. Studies of HTER and post-editing productivity were reviewed to connect quality with real translation-workflow effort [10-12]. Research on multilingual hallucination and professional post-editing standards was included to address reliability and quality assurance [13,14]. Recent shared-task evidence was used to contextualize the rapid evolution of current systems. WMT25 research is especially relevant because it evaluated LLMs and online translation providers on more difficult, document-level test material rather than relying only on easy sentence-level benchmarks. Current commercial documentation was also considered where it helped clarify how providers describe their evaluation methodology; such material was treated as product-specific evidence rather than independent proof of universal superiority. The detailed evidence is discussed later in the paper. Recent comparative studies on literary and medical translation, including studies using professional human translations as a comparator, were also examined to capture language-pair and domain effects [19–23].

The analysis was organized around semantic adequacy, fluency, terminology, context, style, cultural appropriateness, factual fidelity, error severity, post-editing effort, scalability, and accountability. Because this is a review rather than a new benchmark, the Results section synthesizes patterns in the literature and does not report newly collected experimental data.

4. Results

The literature synthesis identifies five broad findings. First, AI-based translation has reached a level of fluency, speed, and scalability that makes it useful for a large range of everyday and professional tasks. Its strongest value is often not the production of a final text but the rapid generation of workable candidates. Second, general-purpose LLMs and specialized translation systems should not be treated as identical technologies. LLMs provide unusually strong interactive control, explanation, and stylistic variation, while specialized systems provide mature translation-oriented infrastructure and can perform strongly on many high-volume language-pair tasks. Performance varies substantially by language direction and domain. Third, translation quality is multidimensional. A system can be fluent but wrong, accurate but stylistically unsuitable, or generally strong while making a small number of critical errors. Consequently, quality assurance should combine source-target comparison, terminology checks, document-level review, and human judgment. Direct comparisons reinforce this pattern: performance changes by language pair and text type, and high sentence-level accuracy can conceal set-level errors [5,19–22].

Fourth, post-editing can create productivity benefits when machine output is sufficiently strong. However, weak output can increase human effort, especially when translators must diagnose and reconstruct the source meaning before editing. The usefulness of post-editing therefore depends on both system quality and workflow design. Fifth, human–AI collaboration provides the most defensible general model. AI is valuable for generation, alternatives, consistency checks, and repetitive translation; humans remain essential for ambiguity resolution, cultural judgment, domain expertise, risk management, and final accountability.

5. Discussion

The central finding of this review is that the debate should move beyond the question of whether AI is 'better than humans.' The comparison is often conceptually flawed because humans and AI systems contribute different forms of capability. A translation engine can process thousands of sentences rapidly; a human translator can interpret why a particular sentence should sound restrained, ironic, diplomatic, or emotionally distant. A large language model can produce multiple formulations; a professional translator can determine which formulation preserves the author's intention. Recent research also suggests that translation systems are converging technologically. Google's WMT2025 submission demonstrates that a commercial translation service can incorporate an LLM, supervised fine-tuning, and reinforcement learning while explicitly controlling a literal-versus-creative trade-off [15]. This suggests that the historical distinction between 'MT engine' and 'LLM' is becoming less stable.

At the same time, the literature warns against equating fluency with reliability. Multilingual hallucination research shows that generative systems can produce unsupported content [13]. Human evaluation studies show that the evaluation methodology itself can change system rankings [9]. WMT25 has responded by using more challenging documents and professional annotation, reinforcing the need for evaluation designs that approximate real translation conditions [17]. Automatic metrics remain useful because they enable rapid and reproducible comparison, but they should be interpreted as instruments rather than verdicts. BLEU is useful for historical and controlled comparisons [7]; COMET provides a stronger semantic signal in many settings [8]. Recent work on automated evaluation also indicates that linguistic diversity and reference availability remain important considerations when comparing evaluation systems [18]. Post-editing further demonstrates why system averages are insufficient. A workflow does not benefit from an excellent average score if the difficult segments are precisely the segments that matter most. Segment-level quality estimation and targeted review can improve the allocation of human attention [10].

The practical implication is a risk-based model. Low-risk content can be translated automatically with light review. Medium-risk content can use AI generation followed by systematic post-editing. High-risk content should receive expert human translation or very intensive human verification, regardless of the model's general benchmark performance.

Practical Recommendations

Organizations implementing AI translation should begin by defining the intended use. The same model may be appropriate for informal comprehension and inappropriate for a legal agreement. Quality criteria should therefore be written before production rather than after an error occurs.

Recommended translation practices emphasize preventing errors rather than addressing them after they occur. Before translation begins, it is important to define the purpose, target audience, subject domain, and acceptable quality standards. The translation system should be selected based on demonstrated performance in the relevant language pair and domain rather than relying on universal rankings. Providing glossaries, proper names, units, style guidelines, and contextual information can further improve accuracy and consistency. AI can also be used to generate alternative translations when variations in wording, tone, or style need to be explored instead of automatically accepting the first output. Targeted quality checks should focus on critical elements such as numbers, dates, names, negation, terminology, and instructions.

Human post-editing remains essential for translations intended for publication, external distribution, or other consequential purposes. Additionally, recurring errors should be documented, and validated terminology and style decisions should be incorporated into future workflows. Translation tasks should also be classified as low-risk, moderate-risk, high-risk, or publication-critical to ensure that human review and resources are allocated according to the potential consequences of errors.

Limitations

This review has several limitations. First, it does not perform a new controlled benchmark of ChatGPT,

Google Translate, and DeepL. System behavior can change rapidly as commercial models are updated, so a static ranking would quickly become outdated.

Second, published benchmarks differ in language pairs, domains, prompts, reference translations, evaluators, and metrics. Results should therefore be interpreted as evidence about particular experimental settings rather than universal properties of a system.

Third, some product-specific evidence comes from vendor publications. Such sources can describe methodology and current product positioning, but independent replication is necessary before treating vendor claims as universal conclusions.

Finally, the review focuses primarily on text translation. Speech translation, multimodal translation, terminology extraction, translation memory integration, and privacy-preserving local deployment deserve separate detailed reviews.

6. Conclusion

Artificial intelligence has transformed machine translation into a highly scalable and increasingly interactive form of language technology. Neural architectures and Transformer-based models created the technical foundation for this development, while large language models have expanded translation into an instruction-driven process involving generation, explanation, comparison, and revision. ChatGPT, Google Translate, and DeepL each provide valuable capabilities, but their strengths should be evaluated according to language pair, domain, text type, and translation purpose. Human translators remain indispensable when meaning, culture, style, accountability, and risk require expert judgment. The most reliable future is therefore a collaborative one. AI can accelerate translation and broaden access to multilingual communication, but professional quality depends on human interpretation, systematic evaluation, and responsible post-editing. The guiding principle should be that AI generates possibilities while humans determine what is acceptable.

Declarations

Ethical Approval Statement: Ethical approval was not required.
Conflict of interest: The author declares no conflict of interest.
Funding: This research received no external funding.

References

1. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR; 2015.
2. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv. 2016;1609.08144.
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30. Red Hook (NY): Curran Associates; 2017. p. 5998-6008.
4. OpenAI. GPT-4 technical report. arXiv. 2023;2303.08774.
5. Zhu W, Liu H, Dong Q, Xu J, Huang S, Kong L, et al. Multilingual machine translation with large language models: empirical results and analysis. Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City: Association for Computational Linguistics; 2024. p. 2765-2781. doi:10.18653/v1/2024.findings-naacl.176.
6. Li J, Zhou H, Huang S, Cheng S, Chen J. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. Trans Assoc Comput Linguist. 2024;12:576-592. doi:10.1162/tacl_a_00655.
7. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics; 2002. p. 311-318. doi:10.3115/1073083.1073135.
8. Rei R, Stewart C, Farinha AC, Lavie A. COMET: a neural framework for MT evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics; 2020. p. 2685-2702. doi:10.18653/v1/2020.emnlp-main.213.

9. Freitag M, Foster G, Grangier D, Ratnakar V, Tan Q, Macherey W. Experts, errors, and context: a large-scale study of human evaluation for machine translation. Trans Assoc Comput Linguist. 2021;9:1460-1474. doi:10.1162/tacl_a_00437.
10. Specia L, Farzindar A. Estimating machine translation post-editing effort with HTER. In: Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User. Denver: Association for Machine Translation in the Americas; 2010. p. 33-43.
11. Gaspari F, Toral A, Naskar SK, Groves D, Way A. Perception vs. reality: measuring machine translation post-editing productivity. In: Proceedings of the 11th Conference of the Association for Machine Translation in the Americas. Vancouver: Association for Machine Translation in the Americas; 2014. p. 60-72.
12. Sanchez-Torron M, Koehn P. Machine translation quality and post-editor productivity. In: Proceedings of the AMTA 2016 MT Researchers' Track. Austin: Association for Machine Translation in the Americas; 2016. p. 16-26.
13. Guerreiro NM, Alves DM, Waldendorf J, Haddow B, Birch A, Colombo P, et al. Hallucinations in large multilingual translation models. Trans Assoc Comput Linguist. 2023;11:1500-1517. doi:10.1162/tacl_a_00615.
14. International Organization for Standardization. ISO 18587:2017. Translation services—post-editing of machine translation output—requirements. Geneva: ISO; 2017.
15. Finkelstein M, Kovacs G, Caswell I, Domhan T, Peter JT, Juraska J, Freitag M, Vilar D. Google Translate's research submission to WMT2025. In: Proceedings of the Tenth Conference on Machine Translation. Suzhou: Association for Computational Linguistics; 2025. p. 723-731. doi:10.18653/v1/2025.wmt-1.48.
16. DeepL. How DeepL evaluates translation quality [Internet]. DeepL; [cited 2026 Oct 18]. Available from: https://www.deepl.com/en/translation-quality-whitepaper.
17. Kocmi T, Artemova E, Avramidis E, Bawden R, Bojar O, Dranch K, et al. Findings of the WMT25 general machine translation shared task: time to stop evaluating on easy test sets. In: Proceedings of the Tenth Conference on Machine Translation. Suzhou: Association for Computational Linguistics; 2025.
18. Lavie A, Hanneman G, Agrawal S, Kanojia D, Lo CK, Zouhar V, et al. Findings of the WMT25 shared task on automated translation evaluation systems: linguistic diversity is challenging and references still help. In: Proceedings of the Tenth Conference on Machine Translation. Suzhou: Association for Computational Linguistics; 2025. p. 436-483.
19. Gao R, Lin Y, Zhao N, Cai ZG. Machine translation of Chinese classical poetry: a comparison among ChatGPT, Google Translate, and DeepL Translator. Humanit Soc Sci Commun. 2024;11:835. doi:10.1057/s41599-024-03363-0.
20. Block P, Schaefer J, Maurer F, Storf H, et al. Quality of machine translations in medical texts: an analysis based on standardised evaluation metrics. Stud Health Technol Inform. 2025. doi:10.3233/SHTI251380.
21. Brewster RCL, Gonzalez L, Khazanchi R, Butler S, Chu J, Selcer S, et al. Performance of ChatGPT and Google Translate for pediatric discharge instruction translation. Pediatrics. 2024;154(1):e2023065573. doi:10.1542/peds.2023-065573.
22. Kong M, Fernandez A, Bains J, Milisavljevic A, Brooks KC, Shanmugam A, et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. BMJ Qual Saf. 2026;35(3):150-158. doi:10.1136/bmjqs-2024-018384.
23. Semenov K, Huang X, Zouhar V, Berger N, Zhu D, Oncevay A, Chen P. Findings of the WMT25 terminology translation task: terminology is useful especially for good MTs. In: Proceedings of the Tenth Conference on Machine Translation. Suzhou: Association for Computational Linguistics; 2025. p. 554-576. doi:10.18653/v1/2025.wmt-1.30.

Disclaimer: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.