RESEARCH STUDY

Cross-Border AI Localisation, Cultural Fidelity and the Continuing Role of Human Expertise

Integrated European and International Evidence Synthesis, 2016–13 September 2026

Marianna Csarnai
Associate Professor; Programme Director, Executive MBA in Digital Marketing Budapest Metropolitan University (**METU**), Budapest, Hungary

REFERENCING
American Psychological Association, 7th edition (author–date)

2026

Executive summary

This integrated study brings together the theoretical, empirical and governance evidence on **cross-border AI localisation, cultural fidelity and human correction**, combining the earlier evidence base with fresh European and international research available through 13 September 2026. Its central question is not whether contemporary LLMs can translate fluent text—they often can—but whether an automated output can reliably perform the communicative, cultural and professional function expected of a target-language text. Earlier translation-evaluation traditions built around string overlap and edit distance remain useful for narrow technical questions, but they do not by themselves measure humour, register, idiomaticity, social meaning, brand fit or the appropriateness of cultural adaptation (Appen, n.d.; Devatine & Abraham, 2024; Van Doren & Holland, 2025).

The strongest operational finding is that the simple opposition **"AI versus human translation" is becoming obsolete**. The evidence increasingly favours structured hybrid systems in which translation memories, terminology, style specifications, domain examples and document context guide an MT or LLM draft; automated quality estimation or AI post-editing catches predictable errors; and professional linguists concentrate on semantic, terminological, pragmatic and cultural decisions. A 2026 production study covering **71,262 segments, ten target languages, five domains and five systems** found that AI post-editing informed by language-specific style guides and high-quality bilingual examples consistently outperformed generic translation baselines. Its human evaluation covered **6,618 segments assessed by 60 professional translators**, while direct LLM translation remained weaker (Nunziatini & Speroni, 2026).

At the same time, cultural localisation remains measurably harder than grammatical translation. In the 2026 GEM cultural-nuance benchmark, seven multilingual LLMs were evaluated in 15 target languages by five native-speaker raters per language. Mean full-text quality was **1.68/3**; category means were **2.20 for holidays, 2.19 for cultural concepts, 1.65 for idioms and 1.45 for puns**. This difference is consequential: a translation may preserve denotative information while failing to reproduce humour, persuasion, identity, politeness or rhetorical effect (Van Doren et al., 2026).

Human intervention is therefore best understood as a **risk-allocation mechanism rather than a universal per-word requirement**. Post-editing can increase throughput, but gains vary by language, domain and system, and highly fluent AI output can increase the need for vigilance because subtle semantic or cultural errors are harder to detect. Controlled professional studies show that post-editing is often faster than translation from scratch, while cognitive and evaluative effort does not disappear and may shift from generation to monitoring (Sarti et al., 2022; Sarti et al., 2025; Terribile, 2023; Wang et al., 2026).

Language-resource inequality remains structural. UNESCO reports more than 7,000 languages globally but only around 1,000 with meaningful online representation. Low-resource and under-represented languages remain disproportionately exposed to weak tokenisation, sparse native data, English-centric benchmark design and translationese. Yet the newer evidence also corrects a deterministic view: targeted corpora, related-language transfer, discourse-level context, dictionary or linguistic scaffolding and fine-tuning can materially narrow the gap (Chen et al., 2024; Compaore et al., 2026; Guo et al., 2024; Lu et al., 2024; Pashchenko et al., 2025; Purason et al., 2024; UNESCO, 2025; Yankovskaya et al., 2023).

Finally, quality standards and regulation require precise interpretation. **ISO 18587:2017 defines requirements for full human post-editing when that service is performed; it is not**

a general law requiring every machine translation to be human-reviewed. ISO 5060:2024 provides a configurable framework for analytic human evaluation of translation output, while the 2026 draft revision of ISO 18587 broadens the terminology to "non-human translation output." In the EU, Article 50 of the AI Act introduces transparency obligations for certain AI-generated content, but it likewise does not create a blanket human-post-editing mandate for translation (European Commission, 2026b; European Parliament and Council of the European Union, 2024; International Organization for Standardization, 2017; International Organization for Standardization, 2024; International Organization for Standardization, 2026).

The resulting production model is **risk-adaptive localisation**: automate predictable linguistic transfer; ground generation in specifications and validated language assets; use QE and AI post-editing for triage; and reserve accountable local human expertise for content where meaning depends on context, culture, consequence and communicative purpose.

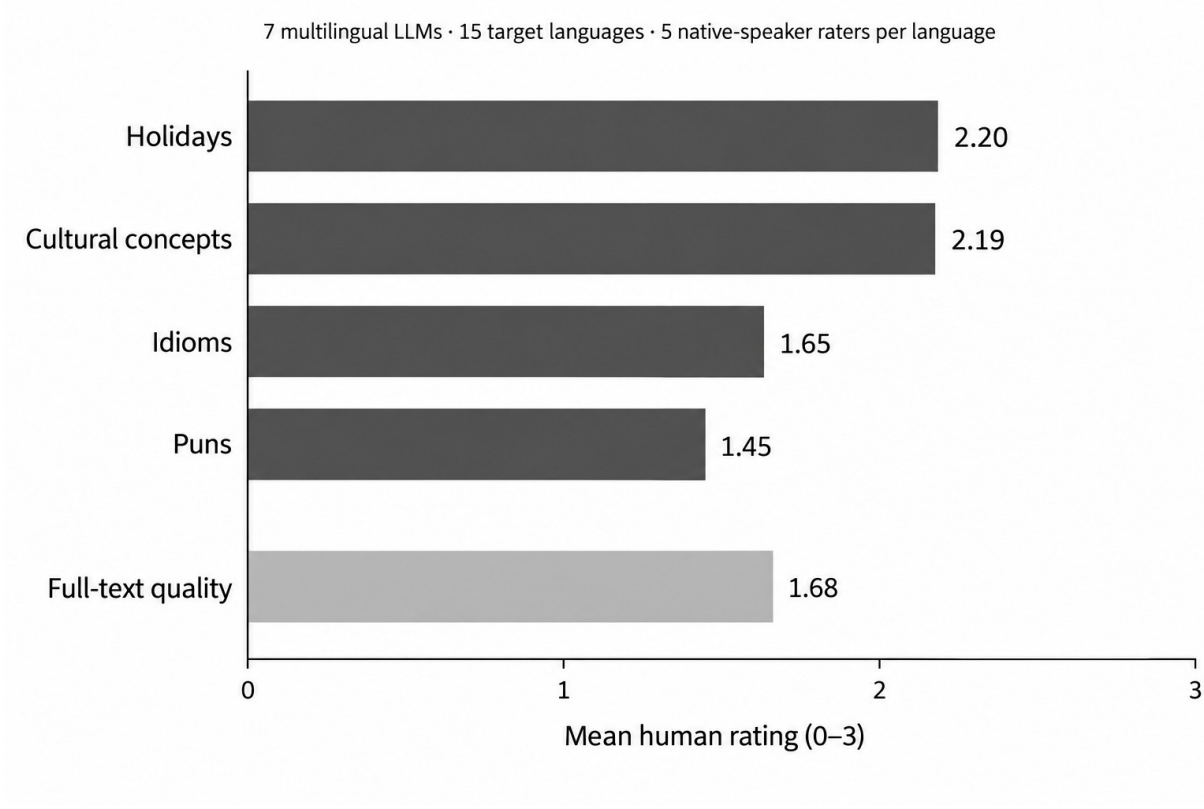


Figure 1. Human-rated quality of culturally nuanced LLM translation. Category means reconstructed from the 2026 GEM benchmark. Source: (Van Doren et al., 2026).

Scope, research questions and evidence methodology

The integrated analysis addresses four linked questions: how close current LLM-based translation systems are to production-level localisation; which error classes remain resistant to scale, prompting and automated post-editing; how language-resource level, domain and cultural density moderate outcomes; and what quality-governance architecture is justified for European and international organisations.

The evidence base spans translation studies, machine translation, multilingual NLP, post-editing research, standards and public-sector deployment. Priority is given to peer-reviewed journals, ACL/EAMT/EMNLP/NAACL/LREC proceedings, official ISO materials, European Commission and EUR-Lex sources, and international institutional evidence. Preprints and industry

benchmarks are retained where they contributed to the earlier evidence base, but they are treated as **lower-certainty or time-sensitive evidence** rather than as interchangeable with peer-reviewed research. The original source register is retained as references **1–56** to preserve traceability; additional European and international sources incorporated in the integrated analysis are numbered **57 onward**.

This study does **not** present a new statistical meta-analysis. The literature mixes ordinal human ratings, BLEU/chrF/COMET scores, HTER and other edit distances, words-per-hour, eye-tracking variables, MQM-like error counts, preference judgements and domain-specific outcomes. A valid pooled Hedges' *g* would require a harmonised extraction dataset with study-level sample sizes, means or effect estimates, variances and comparable control conditions. PRISMA 2020 requires transparent identification, selection, appraisal and synthesis; methodological discussion of effect-size conversion alone is not sufficient to establish a reproducible meta-analysis (Brini & Leung, 2026; Mitchell, 2026; Page, McKenzie, et al., 2021; Page, Moher, et al., 2021; Taherinezhad et al., 2026).

The appropriate methodological stance is therefore a **structured evidence synthesis with quantitative anchors**: numerical results are reported where primary studies support them, while cross-study conclusions are expressed conditionally and moderators are made explicit. This avoids false precision while still allowing a strong operational conclusion about when automation is dependable and when human cultural expertise remains necessary.

Conceptual foundations: from linguistic equivalence to cultural function

A central distinction in localisation is the difference between transferring propositional content and achieving the intended **function of the target text**. Skopos theory treats the purpose of the target text within its target culture as the primary determinant of translation strategy. In creative and persuasive content, literal fidelity may therefore be less important than reproducing the intended effect: a joke must still amuse, a slogan must still persuade, a user-interface warning must still trigger the right action, and a brand voice must remain recognisable (Bal & Köktürk, 2025; Kumar, 2024; Reiß & Vermeer, 2014; "Introducing Translation Studies", n.d.).

This functional perspective explains why apparently fluent LLM output can still be inadequate. Werner Koller's equivalence typology distinguishes denotative, connotative, text-normative, pragmatic and formal-aesthetic equivalence; Venuti's domestication/foreignisation distinction and House's overt/covert translation distinction similarly show that cultural markers can be handled in legitimately different ways depending on purpose. A target text can be grammatically excellent yet pragmatically wrong if it chooses the wrong register, removes a culturally meaningful reference, translates an idiom literally or preserves wordplay that no longer works (TranslationAI, n.d.).

The useful implication for AI is not that models are incapable of adaptation. Rather, **adaptation requires an explicit specification**. Without clear information about audience, locale, register, genre, brand voice, cultural policy and communicative purpose, a general-purpose model has no stable criterion for deciding whether to preserve, explain, replace or recreate a source-culture element. Specification-aware prompting and retrieval of validated examples can close part of this gap, but the target-culture decision itself remains a governance choice rather than a purely linguistic optimisation problem (Hu et al., 2025; Kayano & Sugawara, 2025; Nunziatini & Speroni, 2026; Strandvik, 2025).

Table 1. Functional dimensions of cultural fidelity, synthesising translation-theory frameworks cited in (Bal & Köktürk, 2025; Kumar, 2024; Reiß & Vermeer, 2014; Siegel & The Scale Research Team, 2025; TranslationAI, n.d.; Translators' Journal, 2025; "Introducing Translation Studies", n.d.; "Skopos Theory", n.d.).

Dimension	What must be preserved or adapted	Why fluent AI can still fail
Denotative	Facts, entities, propositional meaning	Hallucination, omission or semantic drift can survive smooth grammar.
Connotative / register	Tone, politeness, social positioning, brand voice	The output may be accurate but socially or stylistically inappropriate.
Text-normative	Genre conventions and expected structure	A literal translation may violate target-market conventions.
Pragmatic	Reader effect, call to action, humour, persuasion	Communicative purpose can fail even when sentence meaning is recoverable.
Formal-aesthetic	Wordplay, rhythm, sound pattern, rhetorical form	Puns, slogans and literary effects often require recreation rather than transfer.

Quality measurement: beyond surface similarity

Older MT evaluation relied heavily on n-gram or string-overlap metrics such as BLEU, ROUGE and METEOR. These can be useful for system development but are poorly aligned with the core questions of localisation because they do not directly measure pragmatic resonance, cultural appropriateness, brand fit or whether an alternative translation is functionally better despite being lexically different (Devatine & Abraham, 2024).

Post-editing research adds a second family of measures: temporal effort, technical effort and cognitive effort. Words-per-hour gives a direct productivity signal; HTER or TER measures insertions, deletions, substitutions and shifts; eye tracking and related measures estimate cognitive engagement. The LLM era exposes limitations in simple edit-distance metrics because a fluent output may need a deep conceptual rewrite rather than many local character edits. Work on compression-based edit distances has therefore been proposed to better represent relocations, duplications and structural changes (Devatine & Abraham, 2024; Minder et al., 2026; O'Brien, 2011).

For professional quality assurance, **analytic error typologies** are more informative. MQM-style evaluation distinguishes error categories such as accuracy, terminology, fluency, style and locale convention and then assigns severity. ISO 5060:2024 similarly provides general guidance for analytic human evaluation using error types and penalty points. However, the commonly used illustrative weights—such as 1 for minor, 5–10 for major and 100 for critical—should not be treated as a universal ISO-mandated scale; the standard is configurable and should be adapted to project requirements (International Organization for Standardization, 2024; Lommel et al., 2024; Macri, 2026; Park & Padó, 2024; Smartling, 2026; Soloviev, 2016).

This distinction is important for cultural fidelity. Minor grammatical defects are visible and usually inexpensive to fix. Major pragmatic, terminological or cultural defects can be less frequent but much more consequential. A single mistranslated safety instruction, offensive cultural substitution or brand-damaging slogan can dominate hundreds of otherwise fluent segments.

Table 2. Practical MQM-style severity interpretation. The weights used in individual programmes should be defined by the evaluation specification rather than assumed to be universal ISO values.

Severity	Illustrative operational interpretation	Typical cultural/localisation impact
Minor	Small fluency, punctuation or style deviation; meaning remains usable.	Low immediate risk, though repeated minor errors can produce "translationese".
Major	Meaning, terminology, register, idiom or required style is materially wrong.	High impact on trust, comprehension, persuasion or brand consistency.
Critical	Error creates safety, legal, reputational or serious cultural harm.	Potential publication blocker; requires escalation and accountable human decision.

Human correction, post-editing effort and productivity

The economic attraction of machine translation and LLM drafting is straightforward: a useful first pass can reduce the amount of text a professional must generate from scratch. Earlier professional estimates often place unaided human translation around **250–500 words per hour**, while some NMT post-editing studies reported approximately **495–934 words per hour** for particular high-resource European language pairs. These ranges should be treated as contextual observations, not universal rates, because domain difficulty, MT quality, translator expertise, interface and acceptance criteria strongly moderate performance (Terribile, 2023).

Controlled evidence supports the general productivity benefit while rejecting a single global multiplier. The **DivEMT** experiment used 18 professional translators working from English into Arabic, Dutch, Italian, Turkish, Ukrainian and Vietnamese; post-editing was consistently faster than translation from scratch, but the size of the gain varied materially by language and system (Sarti et al., 2022). A 2025 literary-translation study likewise found significantly lower editing times when professionals post-edited LLM-generated text, while preserving a similar measured level of creativity (Castaldo et al., 2025).

Generative flexibility also creates new costs. A study of literary automatic post-editing found that ChatGPT changed more text than professional translators, corrected fewer MT errors in some conditions and introduced additional problems even while improving lexical richness (Macken, 2024). In a legal-domain study involving **11 professional English–Spanish translators and 120,102 words**, pre-task perceptions and trust in MT were associated with productivity and quality, showing that the human–computer interaction is part of the system rather than an external afterthought (Briva-Iglesias & O'Brien, 2024). A 2026 German-to-Spanish economic-translation study found terminology to be the largest error category at **34.1%**, while fluent output could encourage trust in subtle semantic errors and post-editors themselves could introduce errors through unnecessary over-editing (Mihalache & Varela Salinas, 2026).

The result is a characteristic **post-editing paradox**. Physical keystrokes and elapsed time may fall while evaluative vigilance rises. Eye-tracking and cognitive-effort research shows that post-editors must monitor a text that often looks convincing even when it contains a subtle error. The cognitive task therefore shifts from generation toward verification: terminology checking, source comparison, hallucination detection, consistency management and cultural judgement (Minder et al., 2026; Wang et al., 2026).

Production evidence from 2026 strengthens the hybrid case. In the 71,262-segment study, AI post-editing informed by language-specific style guides and bilingual examples outperformed generic baselines across a broad production sample, while direct LLM translation did not reach the same quality (Nunziatini & Speroni, 2026). The economic conclusion is not that human work disappears, but that **human labour moves upstream into specification and downstream into high-value review**.

Comparative evidence from major recent studies

Table 3. Comparative evidence. Sources: (Nunziatini & Speroni, 2026; Sarti et al., 2022; Sarti et al., 2025; Thellmann et al., 2026a; Uguet et al., 2024; Van Doren et al., 2026; Velazquez et al., 2025).

Study / dataset	Scale and design	Principal result	Evidence implication
DivEMT, 2022	18 professionals; 6 typologically diverse target languages	Post-editing faster than translation from scratch; gains varied by language/system.	Productivity benefits are real but not language-neutral.
LLM post-translation study, 2024	GPT-4, Claude 3, Gemini; 4 languages; pharmaceutical domain	Potential in automatic post-editing and MQM-style analysis; task/model dependence remained.	AI can be a second-stage quality tool, not only a translator.
LangMark, 2025	206,983 human-annotated source/MT/post-edit triplets; 7 target languages	Few-shot LLMs improved strong MT outputs through automatic post-editing.	High-quality post-edit data is a strategic asset.
QE4PE, 2025	42 professional post-editors; 2 directions	QE highlights could assist post-editing; usefulness varied by domain, language and editor.	Prediction accuracy is not identical to workflow value.
GEM cultural benchmark, 2026	7 LLMs; 15 target languages; 5 native raters per language	Overall 1.68/3; idioms 1.65; puns 1.45.	Cultural/pragmatic fidelity remains below surface fluency.
Production AIPE, 2026	71,262 segments; 10 languages; 5 domains; 6,618 human-rated segments	Style-guide/example retrieval + AI post-editing beat generic baselines; direct LLMT weaker.	Strong evidence for controlled specification-aware hybrid production.
Benchmark-error study, 2026	Expert annotations + MQM-style LLM/QE comparison	Translation errors measurably reduced downstream multilingual evaluation accuracy.	Evaluation data itself requires linguistic and cultural QA.

Cultural nuance, idioms, humour and transcreation

The clearest frontier is **culturally dense language**. The MultiNRC benchmark was designed natively in languages such as French, Spanish and Chinese to test language-specific linguistic reasoning, cultural traditions and wordplay rather than merely translating English-authored questions. Its construction deliberately selected difficult items that multiple state-of-the-art systems failed, exposing a difference between multilingual surface competence and culturally grounded reasoning (Siegel & The Scale Research Team, 2025).

A pilot study of **87 LLM-generated translations of nuanced e-commerce marketing emails across 20 languages** likewise found that figurative expressions, idioms and puns remained recurrent failure points even when grammar and token-level accuracy were strong (Appen, n.d.; Van Doren & Holland, 2025). The later GEM benchmark made the category effect quantitatively visible: holidays and explicit cultural concepts were substantially easier than idioms and puns (Van Doren et al., 2026).

These failures are functional rather than cosmetic. If the communicative purpose of an advertisement is humour, a literal rendering that destroys the joke is not a slightly worse translation; it fails the *Skopos* of the text. The same applies to slogans, idioms, politeness strategies, culturally embedded metaphors and narrative voice. Human transcreation remains particularly valuable because the task is often to invent a **new target-language solution with equivalent rhetorical force**, not to recover a single pre-existing answer (Kumar, 2024; Van Doren & Holland, 2025).

More recent cultural-awareness benchmarks broaden the picture beyond translation. **GIMMICK** evaluated 20 large vision-language models and 11 LLMs across 728 cultural events or facets and reported strong Western-cultural biases, better knowledge of tangible cultural objects than intangible practices, and difficulty moving from broad cultural identification to nuanced understanding (Schneider et al., 2025). **MAKIEval** evaluated seven LLMs across 13 languages, 19 countries or regions and six cultural topics and found that cultural awareness could be stronger when prompting in English—evidence that the language used to access a model can change what cultural knowledge becomes available (Zhao et al., 2025).

The practical conclusion is not that AI has no role in transcreation. LLMs can generate alternatives, explain source-language wordplay, explore rhyme or register options and speed ideation. The accountable final decision, however, should remain with a target-culture specialist when the content carries significant brand, social or cultural consequence.

Low-resource and under-represented languages

Language-resource inequality is one of the most persistent moderators of multilingual AI quality. UNESCO reports more than **7,000 languages globally but only around 1,000 represented meaningfully online**, creating an immediate asymmetry in the data available for pretraining, instruction tuning, evaluation and terminology development (UNESCO, 2025). Earlier work on Irish and extremely low-resource Finno-Ugric languages likewise documents the difficulty of obtaining production-level performance from generic multilingual models without targeted adaptation (Purason et al., 2024; Tran et al., 2024; Zhong et al., 2024).

European evidence shows both the weakness of generic zero-shot systems and the value of intervention. Work across **20 low-resource Finno-Ugric languages** expanded corpora and added manually translated FLORES evaluation sets for nine languages, demonstrating gains from carefully selected back-translation strategies (Yankovskaya et al., 2023). Follow-up work on Proper Karelian, Livvi, Ludian and Veps found that **paragraph-level models** could improve discourse handling while maintaining sentence-level quality (Pashchenko et al., 2025). A Catalan-to-Chinese case found that multilingual training with Spanish as a related higher-resource bridge improved both quality and lexical diversity over baseline approaches (Chen et al., 2024).

Prompting can also compensate for weak exposure. The TALENT “textbook prompting” approach evaluated **112 low-resource languages** and reported an average **14.8%**

improvement over zero-shot baselines across three metrics (Guo et al., 2024). Chain-of-Dictionary prompting produced large gains for some FLORES-200 directions by supplying external lexical knowledge (Lu et al., 2024). In 2026, a French–Mooré study built approximately **83,000 validated sentence pairs** and found a fine-tuned NLLB system to outperform the alternatives tested, underscoring that native data can remain more valuable than model scale alone (Compaore et al., 2026).

A second problem is evaluation validity. Machine-translated benchmarks can import source-side assumptions and target-side errors. Thellmann and colleagues showed that translation errors in multilingual benchmarks are systematically associated with percentage-point losses in measured LLM accuracy and that automatic MQM-style span detection does not trivially reproduce expert human annotation (Thellmann et al., 2026a). SIB-200 and related multilingual evaluations likewise show persistent high-/low-resource gaps (Adelani et al., 2024). A 2026 socio-cultural localisation study of mathematical word problems showed that substituting local names, organisations, currencies and contexts can reveal differences hidden by simple translated English-centric tasks (Azime et al., 2026).

The evidence therefore supports a **conditional rather than deterministic conclusion**: generic zero-shot systems are least reliable where native data, culturally grounded evaluation and linguistic infrastructure are weak, but targeted corpora, related-language transfer, contextual modelling and linguistic scaffolding can substantially narrow the gap.

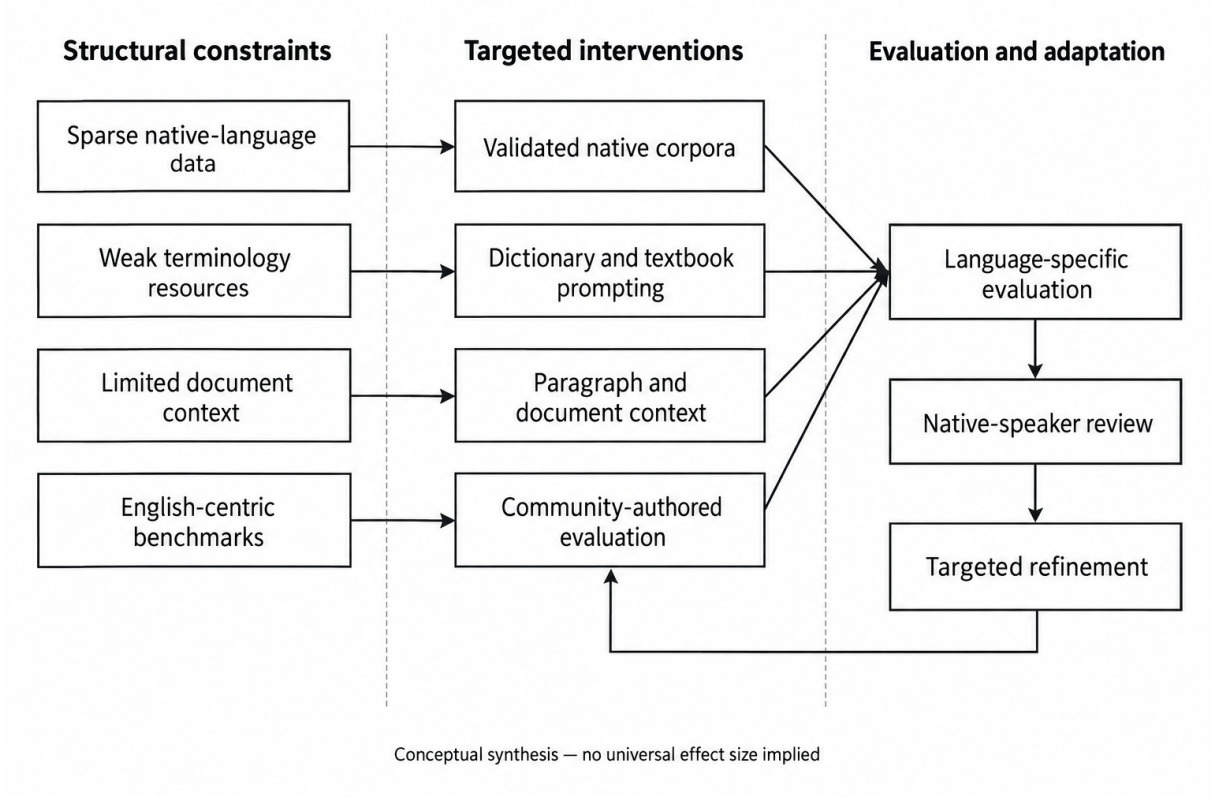


Figure 2. Targeted interventions for low-resource and under-represented languages. Conceptual synthesis of the study's evidence; arrows show workflow relationships, not measured causal effects. Sources: (Chen et al., 2024; Compaore et al., 2026; Guo et al., 2024; Lu et al., 2024; Pashchenko et al., 2025; Purason et al., 2024; UNESCO, 2025; Yankovskaya et al., 2023).

Context, specification-aware prompting and model routing

Traditional MT pipelines largely operate on the source segment plus model-internal knowledge. LLM systems make it possible to place the translation task inside a richer instruction envelope: target audience, locale, communicative purpose, brand voice, terminology, prohibited wording, preferred reference translations, formatting rules and cultural-adaptation policy. This transforms localisation from an underspecified language-generation task into a constrained production task (Kayano & Sugawara, 2025; Strandvik, 2025; TranslationAI, n.d.).

A specification-aware MT framework described in the earlier evidence base incorporated translation specifications into prompting and, in case studies involving Japanese investor-relations texts, reported blind human preferences favouring specification-aware LLM output over proprietary NMT and some official human translations. Because this result comes from a recent preprint rather than a mature multi-site production literature, it should be treated as promising evidence rather than a universal superiority claim (Kayano & Sugawara, 2025).

Independent document-level work points in the same direction. A 2025 study found that **source-primed multi-turn translation** outperformed both whole-document single-pass generation and independently translated segments across several automatic metrics, supporting the intuition that document context changes translation quality (Hu et al., 2025). The 2026 production AI-post-editing study provides the strongest operational confirmation: language-specific style guides and validated bilingual examples improved output over generic translation at large scale (Nunziatini & Speroni, 2026).

Model selection can be part of this routing layer, but vendor rankings age quickly. Industry comparisons have variously associated Claude-family models with stylistic and brand-sensitive copy, GPT-family systems with strong constraint following, Gemini-family systems with long-context workflows and DeepSeek-family systems with cost-efficient high-volume drafting (Khan, 2026; Wolff, 2026). These are **illustrative, version-dependent observations**, not stable scientific rankings. The more durable procurement principle is to benchmark the exact model/version on the organisation's own language pairs, domains, placeholders, terminology and cultural acceptance criteria (Nunziatini & Speroni, 2026; Uguet et al., 2024).

Table 4. Evidence-informed routing factors for localisation workflows.

Routing factor	Lower-risk configuration	Higher-risk configuration
Communicative purpose	Informational gist, routine comprehension	Persuasion, humour, public trust, legal or safety consequence
Context supplied	Source segment only	Document context + glossary + style guide + validated bilingual examples
Cultural density	Literal factual prose	Idioms, slogans, metaphors, culturally embedded references
Language resources	Well-resourced pair with validated benchmark	Under-represented language, dialect or weak native benchmark coverage
Acceptance control	Automated checks / sampling	Qualified human review and local-market sign-off

LLMs as post-editors, judges and quality-estimation systems

LLMs are increasingly used not only to translate but to **criticise, post-edit and score other translations**. GPT-4-based automatic post-editing improved WMT outputs in several directions, although the same work also documented hallucinated edits (Raunak et al., 2023). LangMark, with **206,983 human-annotated source/MT/post-edit triplets across seven target languages**, showed that few-shot LLMs can improve strong MT outputs when supplied with high-quality post-edit examples (Velazquez et al., 2025).

Quality estimation (QE) offers a complementary strategy: predict where a translation is likely to be wrong before a human spends time reviewing it. The QE4PE study involved **42 professional post-editors** and showed that word-level highlights can assist post-editing, but usefulness varies by domain, language and editor (Sarti et al., 2025). This is a critical distinction: a QE model can be statistically accurate yet poorly designed for human decision-making if it distracts, creates false confidence or highlights low-consequence errors while missing high-consequence ones.

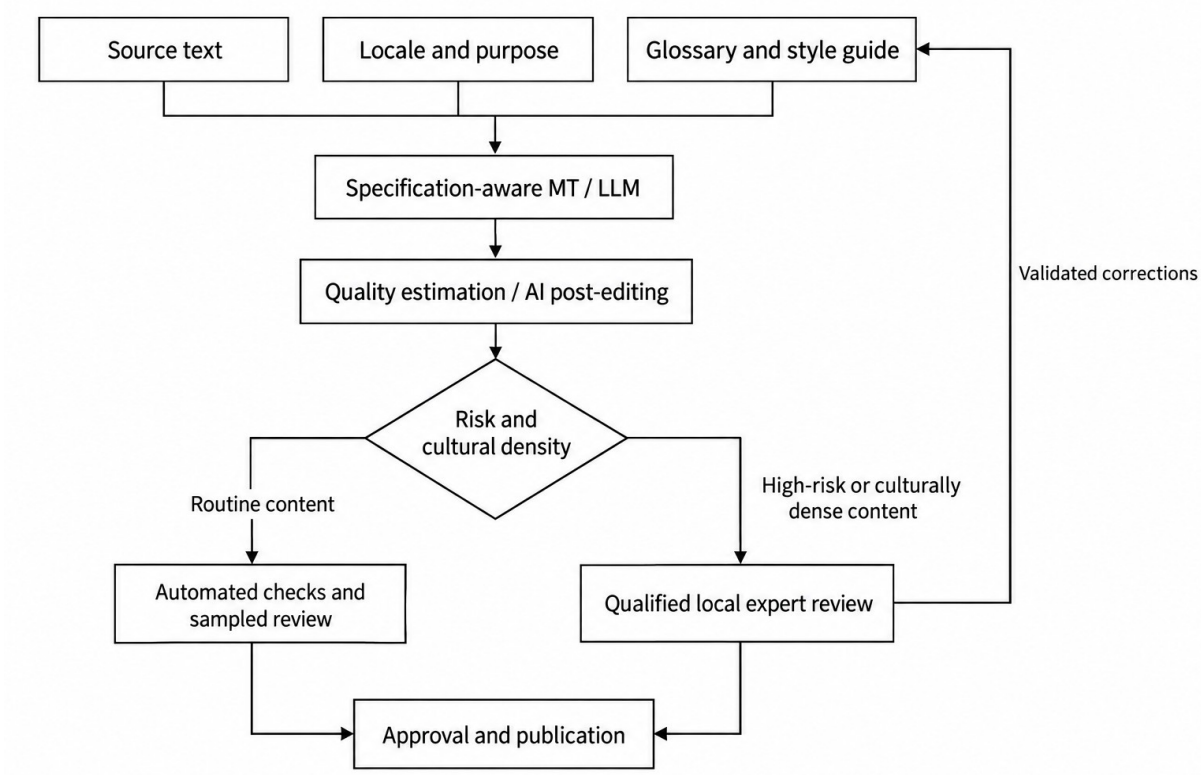


Figure 3. Risk-adaptive localisation workflow synthesised from post-editing, QE, cultural-benchmark and production evidence. Sources: (Nunziatini & Speroni, 2026; Sarti et al., 2022; Sarti et al., 2025; Thellmann et al., 2026a; Van Doren et al., 2026).

LLM-as-judge methods can also produce MQM-like error spans and natural-language rationales. They are useful for triage, regression testing and large-scale sampling, but expert agreement remains challenging on naturally occurring, complex translations (Thellmann et al., 2026a; Zhao et al., 2024). Consequently, automated judging should **route human attention rather than certify cultural acceptability autonomously** in high-risk use cases.

The long-term optimisation target is therefore the **joint human–AI system**. The relevant question is not which component achieves the highest standalone score, but whether the

workflow directs scarce expert attention to the segments where an error is likely, consequential and culturally difficult.

Standards, public-sector practice and EU governance

Standards provide a useful vocabulary for process control, but their scope must be stated accurately. **ISO 18587:2017** specifies requirements for the process of full human post-editing of machine translation output and for post-editor competences. It should not be read as a legal rule that every machine-translated text must be reviewed or as an empirical declaration that machines can never replace a professional translator in any context (ATC Certification, n.d.; European Commission, 2022; International Organization for Standardization, 2017; Interpro Translation Solutions, n.d.; LinguaVox, n.d.; Transpose, 2025). As of September 2026, ISO lists a draft revision—**ISO/DIS 18587**—using the broader phrase "post-editing of non-human translation output," reflecting the shift from classic MT toward generative systems (International Organization for Standardization, 2026).

ISO 5060:2024 addresses translation-output evaluation rather than post-editing process. It supports an analytic approach in which error types and penalty points are configured to create error scores and quality ratings (International Organization for Standardization, 2024; Macri, 2026; oneword, 2024). Organisations can map MQM-style categories to this structure, but should define the error taxonomy, severity rules, thresholds and acceptance criteria explicitly for the project rather than treating one internet-published weighting scheme as the standard itself.

The European Commission's **eTranslation** offers a useful real-world operational distinction. It supports all 24 official EU languages and several non-EU languages, but the Commission describes the output as raw machine translation suitable for gist or as a starting point for human-quality translation. It explicitly warns that professionally accurate high-quality output requires revision by a skilled translator and that machine translation should not be used as the authoritative translation of EU legislation (European Commission, n.d.-a; European Commission, n.d.-b).

The **EU AI Act** adds a separate transparency layer. Article 50 rules apply from 2 August 2026 to certain AI-generated or manipulated content, including specific public-interest text scenarios when content is published without human review or editorial control. These provisions do not create a general human-post-editing obligation for translation, but they make the presence or absence of substantive editorial control relevant to some public-facing AI uses (European Commission, 2026b; European Parliament and Council of the European Union, 2024). The Commission's 2026 Code of Practice on marking and labelling AI-generated content provides a voluntary compliance route for relevant transparency obligations (European Commission, 2026a).

Table 5. Selected governance timeline relevant to AI localisation as of 13 September 2026. Sources: (European Commission, 2026a; European Commission, 2026b; European Parliament and Council of the European Union, 2024; International Organization for Standardization, 2017; International Organization for Standardization, 2024; International Organization for Standardization, 2026).

Date	Instrument / milestone	Practical relevance
2017	ISO 18587	Requirements for full human post-editing of MT output and post-editor competences.

Date	Instrument / milestone	Practical relevance
2024	ISO 5060	General guidance for analytic evaluation of translation output.
2025	EU GPAI obligations begin applying	Broader AI governance context for providers of general-purpose AI models.
Jun–Jul 2026	EU transparency Code + guidance	Implementation support for AI-generated content transparency.
2 Aug 2026	AI Act Article 50 transparency rules apply	Transparency obligations for defined categories of AI-generated/manipulated content.
2026	ISO/DIS 18587	Draft revision broadening scope language to non-human translation output.

Integrated production model: risk-adaptive human oversight

The evidence supports a production architecture in which human review intensity is **proportional to risk, cultural density and communicative purpose**. Applying the same workflow to an internal gist translation and a public campaign slogan is economically inefficient; treating them as equally safe for automation is equally indefensible.

A localisation specification should be created **before generation**. At minimum it should define target locale and audience, communicative purpose, register, terminology, protected names, formatting rules, units and date conventions, brand voice, prohibited formulations, and the policy for domestication, foreignisation or transcreation. Retrieval of validated examples and language-specific style guides should be part of the generation context where available (Hu et al., 2025; Nunziatini & Speroni, 2026).

Culturally dense segments should be evaluated separately from routine prose. Idioms, slogans, humour, metaphors, culturally embedded entities, politeness/register shifts and identity-bearing language should carry stricter thresholds because aggregate quality scores can conceal precisely the categories in which current models perform worst (Van Doren et al., 2026).

Automated QE and LLM-as-judge systems should be used as **triage mechanisms**, not as sole certifiers of cultural quality. Their strongest role is to rank uncertainty, surface probable errors and support regression testing; high-risk acceptance still requires accountable target-language expertise (Sarti et al., 2025; Thellmann et al., 2026a).

Finally, human review itself should be auditable. For consequential publication, the workflow should retain the source, generated draft, relevant terminology/specification version, automated flags, material human edits, reviewer identity or role, and final approval state. This turns “human in the loop” from a vague assurance into a testable production control.

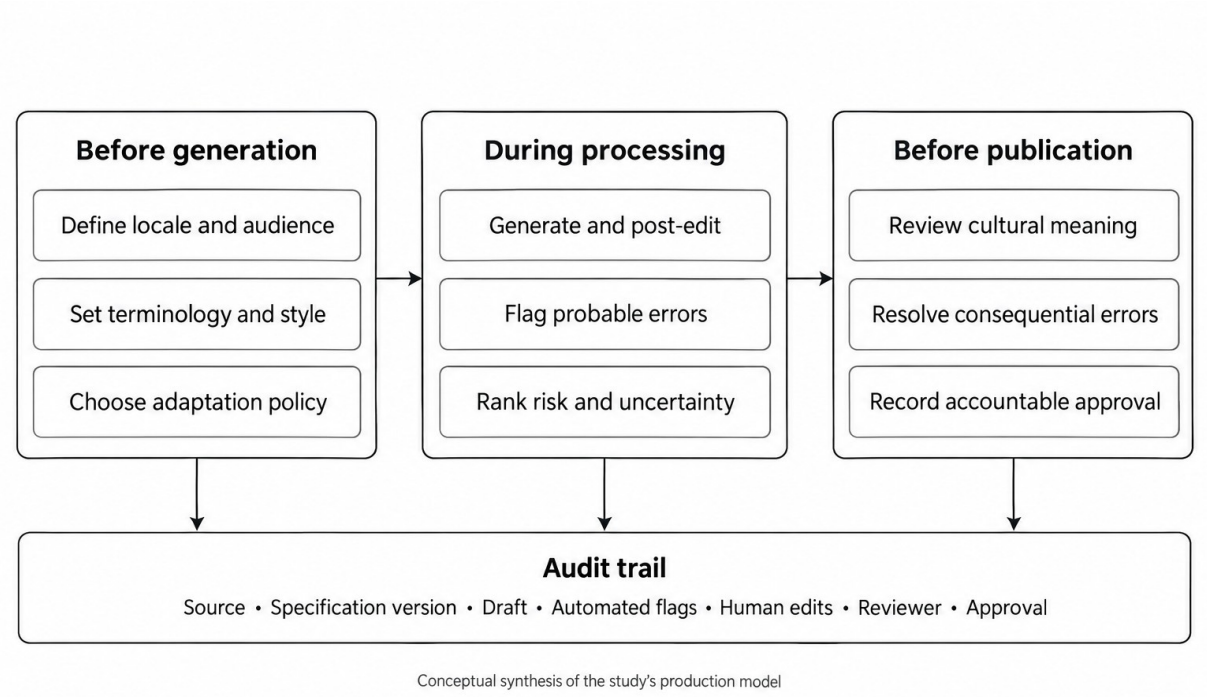


Figure 4. Human responsibilities and the audit trail across the localisation production process. Conceptual synthesis of the integrated production model. Sources: (Hu et al., 2025; Nunziatini & Speroni, 2026; Sarti et al., 2025).

Table 6. Risk-adaptive localisation control model synthesised from the integrated evidence base.

Localisation class	Recommended workflow	Human control	Typical examples
Low-risk / gist	Generic MT or LLM + automated terminology/security checks	Optional or sampled	Internal discovery, low-stakes comprehension, ephemeral material
Routine external	Domain MT/LLM + terminology + style specification + QE/AIPE	Sampled native-speaker QA; escalate flagged items	Help content, product information, routine operations
High-value brand content	Specification-aware LLM + translation memory + cultural brief	Mandatory local-market editor/transcreator	Advertising, campaign copy, slogans, persuasive UX
High-consequence factual	Controlled MT/LLM + approved terminology + traceable source + analytic error review	Full qualified review	Legal, medical, financial or safety-related material
Culturally dense creative work	Human-led transcreation; AI used for alternatives, analysis and research	Human owns final communicative decisions	Humour, puns, literature, culturally sensitive campaigns

Limitations and research agenda

This evidence synthesis has several limitations. It is not a preregistered exhaustive review, and proprietary model behaviour can change between a study, a vendor update and a production deployment. Model-specific league tables therefore have a short shelf life. The most transferable findings concern workflow architecture—context, specification, quality assurance, cultural evaluation and human escalation—rather than any particular vendor.

Human-workflow studies also remain smaller than mainstream machine-learning benchmarks. DivEMT involved 18 professionals and QE4PE 42, while many creative-translation studies are smaller still (Sarti et al., 2022; Sarti et al., 2025). These experiments provide high-value behavioural evidence but should not be generalised into universal productivity coefficients. The 2026 production study materially improves scale with 71,262 segments and 60 professional human evaluators, yet it still represents a particular commercial workflow rather than the entire localisation industry (Nunziatini & Speroni, 2026).

"Cultural fidelity" is not a single universal scalar. Different target communities, regional varieties, age groups, institutions and brands can legitimately prefer different adaptations. Future work therefore requires **native-authored rather than merely translated benchmarks; multiple raters within each locale; regional and dialectal coverage; longitudinal professional-workflow studies; and direct business or societal outcomes** such as comprehension, trust, conversion, complaints and safety incidents (Adelani et al., 2024; Azime et al., 2026; Thellmann et al., 2026a).

A second research priority is **selective human review**: predicting not merely whether a sentence contains an error, but whether the probable error is consequential enough to justify expert intervention. This requires combining QE with severity, domain risk, cultural density and uncertainty (Sarti et al., 2025). A third priority is community-centred low-resource localisation, in which language communities participate in data creation, terminology, evaluation and decisions about acceptable cultural adaptation rather than serving only as passive test sets (UNESCO, 2025).

Conclusion

The transition from statistical MT to neural systems and then to generative LLMs has transformed the economics and ergonomics of localisation. Contemporary models can generate fluent multilingual drafts, exploit broad context, follow detailed specifications and increasingly perform useful automatic post-editing and quality estimation. These advances make large-scale cross-border production possible at a speed and volume that human-only workflows cannot economically match.

They do not, however, eliminate the separate problem of **cultural fidelity**. The strongest international evidence shows a persistent gap between grammatical fluency and pragmatic success, most visibly in idioms, puns, humour, culturally embedded reasoning, specialised terminology and under-represented languages. The gap is not uniform: it shrinks substantially when systems are supplied with project specifications, validated bilingual examples, domain assets, native-language data and document context (Guo et al., 2024; Hu et al., 2025; Lu et al., 2024; Nunziatini & Speroni, 2026; Van Doren et al., 2026).

Accordingly, the defensible model is neither "human-only" nor "AI-only". It is **structured, specification-aware and risk-adaptive localisation**. Machines should handle predictable linguistic transfer and high-volume drafting; automated quality tools should rank and refine; and

accountable local experts should own the decisions in which meaning depends on context, culture, consequence and communicative purpose. That allocation of work is both the strongest reading of the current evidence and the most practical route to scalable cross-border communication without sacrificing cultural fidelity.

References

Adelani, D. I., Liu, H., Shen, X., Vassilyev, N., Alabi, J., Mao, Y., Gao, H., & Lee, E.-S. A. (2024). SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 226–245). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.14

Appen. (n.d.). *Multilingual LLM translation: Evaluating cultural nuance in generative AI.* https://www.appen.com/research/multilingual-cultural-nuance

Arsolon, M. (2025). *Benchmarking LLM translations using semantic similarity, BLEU+TLR, and AI evaluation.* Medium. https://medium.com/@michaelangeloarsolon/benchmarking-llm-translations-using-semantic-similarity-bleu-tlr-and-ai-evaluation-514e9f8adc78

ATC Certification. (n.d.). *ISO 18587 certification for post editing machine translation output.* https://atccertification.com/iso-18587-certification/

Azime, I. A., Belay, T. D., Klakow, D., Slusallek, P., & Chhabra, A. (2026). Bridging the culture gap: A framework for LLM-driven socio-cultural localization of math word problems in low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 855–870). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.42

Bal, D., & Köktürk, Ş. (2025). Evaluating large language models in translation: A theoretical and practical analysis based on Skopos theory. *Söylem*(Çeviribilim Özel Sayısı II), 819–826. https://dergipark.org.tr/en/download/article-file/4443035

Brini, S., & Leung, T. I. (2026). Value and credibility of meta-analysis: Tutorial on enhancing methodological rigor and AI-powered efficiency. *Journal of Medical Internet Research, 28*, e92132. https://doi.org/10.2196/92132

Briva-Iglesias, V., & O'Brien, S. (2024). Pre-task perceptions of MT influence quality and productivity: The importance of better translator-computer interactions and implications for training. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp. 444–454). European Association for Machine Translation. https://aclanthology.org/2024.eamt-1.37/

Castaldo, A., Castilho, S., Moorkens, J., & Monti, J. (2025). *Extending CREAMT: Leveraging large language models for literary translation post-editing.* arXiv. https://arxiv.org/abs/2504.03045

Chen, Y., Toral, A., Li, Z., & Farrús, M. (2024). Improving NMT from a low-resource source language: A use case from Catalan to Chinese via Spanish. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp. 229–245). European Association for Machine Translation. https://aclanthology.org/2024.eamt-1.20/

Compaore, W. S. R., Ouattara, M., Kafando, R., Bissyandé, T. F., Kabore, A. K., & Sabane, A. (2026). Neural machine translation for French–Mooré: Adapting large language models to low-resource languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)* (pp. 615–622). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.loreslm-1.53

Devatine, N., & Abraham, L. (2024). *Assessing human editing effort on LLM-generated texts via compression-based edit distance.* arXiv. https://arxiv.org/abs/2412.17321

European Commission. (2022). *ISO 18587:2017 on post-editing machine translation output – Theory, practice and other related issues [Event programme].* https://commission.europa.eu/system/files/2022-10/tew_programme_27.10.22_en_updated_0.pdf

European Commission. (2026a). *Commission publishes Code of Practice on marking and labelling AI-generated content.* https://digital-strategy.ec.europa.eu/en/news/commission-publishes-code-practice-marking-and-labelling-ai-generated-content

European Commission. (2026b). *Guidelines on transparency obligations for providers and deployers of certain AI systems.* https://digital-strategy.ec.europa.eu/en/policies/guidelines-ai-transparency-obligations

European Commission. (n.d.-a). *AI translation and language tools.* Retrieved September 13, 2026, from https://commission.europa.eu/resources-partners/etranslation_en

European Commission. (n.d.-b). *Use of machine translation on Europa.* Retrieved September 13, 2026, from https://commission.europa.eu/languages-our-websites/use-machine-translation-europa_en

European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act).* Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Feingold, A. (2019). New approaches for estimation of effect sizes and their confidence intervals for treatment effects from randomized controlled trials. *The Quantitative Methods for Psychology, 15*(2), 96–111. https://doi.org/10.20982/tqmp.15.2.p096

Gosling, C. J., Cortese, S., Solmi, M., Haza, B., Vieta, E., Delorme, R., Fusar-Poli, P., & Radua, J. (2025). MetaConvert: An automatic suite for estimation of 11 different effect size measures and flexible conversion across them. *Research Synthesis Methods, 16*(3), 575–586. https://doi.org/10.1017/rsm.2025.11

Guo, P., Ren, Y., Hu, Y., Li, Y., Zhang, J., Zhang, X., & Huang, H.-Y. (2024). Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 15685–15697). Association for Computational Linguistics. https://aclanthology.org/2024.lrec-main.1362/

Hu, H., Vamvas, J., & Sennrich, R. (2025). Source-primed multi-turn conversation helps large language models translate documents. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 23702–23712). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.1289

International Organization for Standardization. (2017). *Translation services — Post-editing of machine translation output — Requirements (ISO Standard No. 18587:2017).* https://www.iso.org/standard/62970.html

International Organization for Standardization. (2024). *Translation services — Evaluation of translation output — General guidance (ISO Standard No. 5060:2024).* https://www.iso.org/standard/80701.html

International Organization for Standardization. (2026). *Translation services — Post-editing of non-human translation output — Requirements (ISO/DIS 18587) [Draft international standard].* https://www.iso.org/standard/88184.html

Interpro Translation Solutions. (n.d.). *What is ISO 18587? A guide to the global standard for machine translation post-editing.* https://www.interproinc.com/what-is-iso-18587-machine-translation-post-editing-guide/

Introducing translation studies: Theories and applications. (n.d.). https://lin2.onecollective.org/manual/T9HIJY703444/Introducing-Translation-Studies-Theories-And-Applications

Kayano, Y., & Sugawara, S. (2025). *Specification-aware machine translation and evaluation for purpose alignment.* arXiv. https://arxiv.org/abs/2509.17559

Khan, S. U. (2026). *Best AI for translators 2026: 10 trusted tools.* Perplexity AI Magazine. https://perplexityaimagazine.com/ai-tools/best-ai-for-translators-2026/

Kumar, M. (2024). Translation and the Skopos theory: A functionalist approach to audiovisual media. *Knowledgeable Research, 3*(2), 60–70. https://knowledgeableresearch.com/index.php/1/article/download/492/430

Lee, S. W., & Koo, M. J. (2022). PRISMA 2020 statement and guidelines for systematic review and meta-analysis articles, and their underlying mathematics: Life cycle committee recommendations. *Life Cycle, 2*. https://doi.org/10.54724/lc.2022.e9

LinguaVox. (n.d.). *ISO 18587: Professional post-editing of machine translation.* https://iso18587.com/

Lommel, A., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., Vaasa, A., Benzo, A., Sparano, R. M., Foresi, M., Innis, J., Han, L., & Nenadic, G. (2024). *The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control.* arXiv. https://arxiv.org/abs/2405.16969

Lu, H., Yang, H., Huang, H., Zhang, D., Lam, W., & Wei, F. (2024). Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 958–976). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.55

Macken, L. (2024). Machine translation meets large language models: Evaluating ChatGPT's ability to automatically post-edit literary texts. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology* (pp. 65–81). Association for Computational Linguistics. https://aclanthology.org/2024.ctt-1.7/

Macri, A. N. (2026). *ISO 5060:2024 explained: A practical guide to translation output evaluation.* LanguageCheck. https://blog.languagecheck.ai/iso-50602024-explained-a-practical-guide-to-translation-output-evaluation/

Mihalache, I., & Varela Salinas, M.-J. (2026). Meaning-making process and error dynamics in ChatGPT-mediated translation. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp. 634–648). European Association for Machine Translation. https://aclanthology.org/2026.eamt-1.40/

Minder, J., Mestivier, A., & Kübler, N. (2026). *Machine translation and post-editing: Comparative evaluation of different MT systems and post-editor groups in specialised translation.* arXiv. https://arxiv.org/abs/2606.23002

Mitchell, S. (2026). *Converting effect sizes: Cohen's d and Hedges g.* Research Gold. https://researchgold.org/blog/converting-between-effect-sizes-or-cohens-d-hedges-g

NexTranslate. (2026). *How translation quality is measured: MQM and benchmarks.* https://www.nextranslate.ai/blog/how-translation-quality-is-measured/

Nunziatini, M., & Speroni, M. (2026). AI post-editing in production: A 71,262-segment evaluation across five domains, ten languages and five systems. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (Volume 2)* (pp. 49–55). European Association for Machine Translation. https://aclanthology.org/2026.eamt-2.24/

oneword. (2024). *New ISO 5060 standard for assessing translation quality.* https://www.oneword.de/en/artikel/new-iso-5060-standard/

Orăsan, C., Mitkov, R., Corpas Pastor, G., & Monti, J. (Eds.). (2023). *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023).* https://hit-it-conference.org/wp-content/uploads/2023/07/HiT-IT-2023-proceedings.pdf

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation, 25*(3), 197–215. https://doi.org/10.1007/s10590-011-9096-7

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ, 372*, n160. https://doi.org/10.1136/bmj.n160

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Park, D., & Padó, S. (2024). *Multi-dimensional machine translation evaluation: Model evaluation and resource for Korean.* arXiv. https://arxiv.org/abs/2403.12666

Pashchenko, D., Yankovskaya, L., & Fishel, M. (2025). Paragraph-level machine translation for low-resource Finno-Ugric languages. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)* (pp. 458–469). Association for Computational Linguistics. https://aclanthology.org/2025.nodalida-1.50/

PRISMA. (n.d.). *PRISMA statement.* Retrieved September 13, 2026, from https://www.prisma-statement.org/

Purason, T., Kuulmets, H.-A., & Fishel, M. (2024). *LLMs for extremely low-resource Finno-Ugric languages.* arXiv. https://arxiv.org/abs/2410.18902

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., & Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 12009–12024). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.804

Reiß, K., & Vermeer, H. J. (2014). *Towards a general theory of translational action: Skopos theory explained.* Routledge. (C. Nord, Trans.). (Original work published 1984) http://traduttologiageneralenz.pbworks.com/w/file/fetch/139448655/Katharina%20Reiss,%20Hans%20J%20Vermeer%20-%20Towards%20a%20General%20Theory%20of%20Translational%20Action_%20Skopos%20Theor.pdf

Sarti, G., Bisazza, A., Guerberof-Arenas, A., & Toral, A. (2022). DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 7795–7816). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.532

Sarti, G., Zouhar, V., Chrupała, G., Guerberof-Arenas, A., Nissim, M., & Bisazza, A. (2025). QE4PE: Word-level quality estimation for human post-editing. *Transactions of the Association for Computational Linguistics, 13*, 1410–1435. https://doi.org/10.1162/tacl.a.46

Schneider, F., Holtermann, C., Biemann, C., & Lauscher, A. (2025). GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 9605–9668). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.500

SciSpace. (n.d.). *Hedges g calculator for effect size (bias-corrected).* Retrieved September 13, 2026, from https://scispace.com/agents/hedges-g-calculator-v3y2uwr8

Scourneau, V. (2025). Impact of automatic post-editing and prompting strategies on the linguistic features of English-to-French translations of editorials. *Revista Tradumàtica, 23*, 65–100. https://doi.org/10.5565/rev/tradumatica.533

Siegel, M., & The Scale Research Team. (2025). *The future is multilingual: Scale's new evaluation benchmark.* Scale AI. https://scale.com/blog/multinrc

Skopos theory. (n.d.). Wikipedia. Retrieved September 13, 2026, from https://en.wikipedia.org/wiki/Skopos_theory

Smartling. (2026). *How translation quality scoring works: MQM explained.* https://www.smartling.com/questions/translation-quality-scoring

Soloviev, K. (2016). *Measuring content quality with error typology: Step by step guide.* ContentQuo. https://www.contentquo.com/blog/measuring-content-quality-with-error-typology-step-by-step-guide

Strandvik, I. (2025). Translation quality and the role of specifications – How standards can help the translation sector today. *Across Languages and Cultures, 26*(S), 5–24. https://doi.org/10.1556/084.2025.01057

Taherinezhad, M., Maier, S., Vitagliano, G., Pierri, F., & Feuerriegel, S. (2026). *AutoSynthesis: An agentic system for automated meta-analysis.* arXiv. https://arxiv.org/abs/2607.15247

Terribile, S. (2023). Is post-editing really faster than human translation?. *Translation Spaces*. Advance online publication. https://doi.org/10.1075/ts.22044.ter

Thellmann, K., Stadler, B., Färber, M., & Lehmann, J. (2026a). Quantifying the impact of translation errors on multilingual LLM evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 41311–41334). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.1916

Thellmann, K.-D., Stadler, B., Färber, M., & Lehmann, J. (2026b). *Quantifying the impact of translation errors on multilingual LLM evaluation.* arXiv. https://arxiv.org/abs/2605.24904

Tommi P. (2025). *Embracing AI in localization: A 2025–2028 roadmap.* Medium. https://medium.com/@hastur/embracing-ai-in-localization-a-2025-2028-roadmap-a5e9c4cd67b0

Tran, K.-T., O'Sullivan, B., & Nguyen, H. D. (2024). *Irish-based large language model with extreme low-resource settings in machine translation.* ACL. https://cora.ucc.ie/items/2164fe40-9610-457a-bb90-50293e2edacf

TranslationAI. (n.d.). *8 translation frameworks explained.* https://translationai.app/theory.html

Translators' Journal. (2025). *Origins and development of Skopos theory: Knowing the 'why' of translation.* https://translatorsjournal.com/origins-and-development-of-skopoS-theory/

Transpose. (2025). *ISO 18587 compliance: Machine translation post-editing requirements.* https://transpose.ch/en/blog/iso-18587-compliance-machine-translation-post-editing-requirements

Uguet, C., Bane, F., Aymo, M., Torres, J., Zaretskaya, A., & Blanch Miró, T. (2024). LLMs in post-translation workflows: Comparing performance in post-editing and error analysis. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp. 373–386). European Association for Machine Translation. https://aclanthology.org/2024.eamt-1.32/

UNESCO. (2025). *UNESCO launches global roadmap on multilingualism in the digital era.* https://www.unesco.org/en/articles/unesco-launches-global-roadmap-multilingualism-digital-era

Van Doren, M., & Holland, C. (2025). *"Be my cheese?": Assessing cultural nuance in multilingual LLM translations.* arXiv. https://arxiv.org/abs/2509.21577

Van Doren, M., Ford, C., Barajas, J., VanMeter, R., & Holland, C. (2026). "Be my cheese?": Cultural nuance benchmarking for machine translation in multilingual LLMs. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)* (pp. 52–76). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.gem-main.6

Velazquez, D., Grace, M., Karageorgos, K., Carin, L., Schliem, A., Zaikis, D., & Wechsler, R. (2025). LangMark: A multilingual dataset for automatic post-editing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 32653–32667). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1569

Wang, F., Xie, H., & Zhang, X. (2026). Association between machine translation post-editing and post-editors' cognition: A three-level meta-analysis based on eye-tracking evidence. *Behavioral Sciences, 16*(3), 365. https://doi.org/10.3390/bs16030365

Wolff, R. (2026). *What is the best LLM for translation? A comparison of top AI translation models.* Lokalise. Retrieved September 13, 2026, from https://lokalise.com/blog/what-is-the-best-llm-for-translation/

Yankovskaya, L., Tars, M., Tättar, A., & Fishel, M. (2023). Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 762–771). Association for Computational Linguistics. https://aclanthology.org/2023.nodalida-1.77/

Zhao, H., Liu, Y., Tao, S., Meng, W., Chen, Y., Geng, X., Su, C., Zhang, M., & Yang, H. (2024). *From handcrafted features to LLMs: A brief survey for machine translation quality estimation.* arXiv. https://doi.org/10.1109/IJCNN60899.2024.10650457

Zhao, R., Chen, B., Plank, B., & Hedderich, M. A. (2025). MAKIEval: A multilingual automatic WiKidata-based framework for cultural awareness evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 23104–23136). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.1256

Zhong, T., Yang, Z., Liu, Z., Zhang, R., You, W., Liu, Y., Sun, H., Pan, Y., Li, Y., Zhou, Y., Jiang, H., Chen, J., Li, X., & Liu, T. (2024). *Opportunities and challenges of large language models for low-resource languages in humanities research.* arXiv. https://arxiv.org/abs/2412.04497

Integrated source register

References 1–56 retain the numbering of the earlier 2026 evidence synthesis so that its source traceability is not lost. New sources incorporated into the integrated European and international analysis begin at reference 57. URLs are preserved in full; model- or industry-specific sources should be interpreted in light of their evidence status and publication date.

1. (Devatine & Abraham, 2024) https://arxiv.org/html/2412.17321v1

2. (Minder et al., 2026) https://arxiv.org/pdf/2606.23002

3. (Arsolon, 2025) https://medium.com/@michaelangeloarsolon/benchmarking-llm-translations-using-semantic-similarity-bleu-tlr-and-ai-evaluation-514e9f8adc78

4. (Van Doren & Holland, 2025) https://arxiv.org/abs/2509.21577

5. (Appen, n.d.) https://www.appen.com/research/multilingual-cultural-nuance

6. (Van Doren et al., 2026) https://aclanthology.org/2026.gem-main.6.pdf

7. (Kumar, 2024) https://knowledgeableresearch.com/index.php/1/article/download/492/430

8. ("Introducing Translation Studies", n.d.) https://lin2.onecollective.org/manual/T9HIJY703444/Introducing-Translation-Studies-Theories-And-Applications

9. (Reiß & Vermeer, 2014) http://traduttologiageneralenz.pbworks.com/w/file/fetch/139448655/Katharina%20Reiss,%20Hans%20J%20Vermeer%20-%20Towards%20a%20General%20Theory%20of%20Translational%20Action_%20Skopos%20Theor.pdf

10. ("Skopos Theory", n.d.) https://en.wikipedia.org/wiki/Skopos_theory

11. (Translators' Journal, 2025) https://translatorsjournal.com/origins-and-development-of-skopoS-theory/

12. (Bal & Köktürk, 2025) https://dergipark.org.tr/en/download/article-file/4443035

13. (Siegel & The Scale Research Team, 2025) https://scale.com/blog/multinrc

14. (TranslationAI, n.d.) https://translationai.app/theory.html

15. (Kayano & Sugawara, 2025) https://arxiv.org/html/2509.17559v1

16. (Page, Moher, et al., 2021) https://pmc.ncbi.nlm.nih.gov/articles/PMC8005925/

17. (PRISMA, n.d.) https://www.prisma-statement.org/

18. (Page, McKenzie, et al., 2021) https://www.bmj.com/content/372/bmj.n71

19. (Wang et al., 2026) https://www.researchgate.net/publication/401543945_Association_Between_Machine_Translation_Post-Editing_and_Post-Editors'_Cognition_A_Three-Level_Meta-Analysis_Based_on_Eye-Tracking_Evidence

20. (O'Brien, 2011) https://doras.dcu.ie/17154/1/Towards_Predicting_Postediting_Productivity_Final_2.pdf

21. (Mitchell, 2026) https://researchgold.org/blog/converting-between-effect-sizes-or-cohens-d-hedges-g

22. (Taherinezhad et al., 2026) https://arxiv.org/html/2607.15247v1

23. (Gosling et al., 2025) https://www.cambridge.org/core/journals/research-synthesis-methods/article/metaconvert-an-automatic-suite-for-estimation-of-11-different-effect-size-measures-and-flexible-conversion-across-them/A7EBE41C2F4D5EFFBC44F4981D226CE6

24. (Feingold, 2019) https://www.tqmp.org/RegularArticles/vol15-2/p096/p096.pdf

25. (Lee & Koo, 2022) https://www.elifecycle.org/archive/view_article?pid=lc-2-0-9

26. (SciSpace, n.d.) https://scispace.com/agents/hedges-g-calculator-v3y2uwr8

27. (Scourneau, 2025) https://revistes.uab.cat/tradumatica/article/download/n23-scourneau/n23-scourneau/3514

28. (Brini & Leung, 2026) https://www.jmir.org/2026/1/e92132

29. (LinguaVox, n.d.) https://iso18587.com/

30. (ATC Certification, n.d.) https://atccertification.com/iso-18587-certification/

31. (Interpro Translation Solutions, n.d.) https://www.interproinc.com/what-is-iso-18587-machine-translation-post-editing-guide/

32. (International Organization for Standardization, 2017) https://plaintalks.com/uploads/short-url/xAYK1ZRG3LDB6GL2O4HrfU6Nxkt.pdf

33. (European Commission, 2022) https://commission.europa.eu/system/files/2022-10/tew_programme_27.10.22_en_updated_0.pdf

34. (Transpose, 2025) https://transpose.ch/en/blog/iso-18587-compliance-machine-translation-post-editing-requirements

35. (Macri, 2026) https://blog.languagecheck.ai/iso-50602024-explained-a-practical-guide-to-translation-output-evaluation/

36. (oneword, 2024) https://www.oneword.de/en/artikel/new-iso-5060-standard/

37. (NexTranslate, 2026) https://www.nextranslate.ai/blog/how-translation-quality-is-measured/

38. (Lommel et al., 2024) https://arxiv.org/html/2405.16969v5

39. (Orăsan et al., 2023) https://hit-it-conference.org/wp-content/uploads/2023/07/HiT-IT-2023-proceedings.pdf

40. (Soloviev, 2016) https://www.contentquo.com/blog/measuring-content-quality-with-error-typology-step-by-step-guide

41. (Park & Padó, 2024) https://arxiv.org/html/2403.12666v1

42. (Smartling, 2026) https://www.smartling.com/questions/translation-quality-scoring

43. (Strandvik, 2025) https://www.akjournals.com/view/journals/084/26/S/article-p5.xml

44. (Terribile, 2023) https://www.researchgate.net/publication/376643636_Is_post-editing_really_faster_than_human_translation

45. (Tommi P, 2025) https://medium.com/@hastur/embracing-ai-in-localization-a-2025-2028-roadmap-a5e9c4cd67b0

46. (Castaldo et al., 2025) https://arxiv.org/pdf/2504.03045

47. (Castaldo et al., 2025) https://arxiv.org/html/2504.03045v1

48. (Tran et al., 2024) https://cora.ucc.ie/items/2164fe40-9610-457a-bb90-50293e2edacf

49. (Purason et al., 2024) https://arxiv.org/html/2410.18902v1

50. (Zhong et al., 2024) https://arxiv.org/html/2412.04497v5

51. (Thellmann et al., 2026a) https://aclanthology.org/2026.acl-long.1916.pdf

52. (Thellmann et al., 2026b) https://arxiv.org/html/2605.24904v1

53. (Uguet et al., 2024) https://aclanthology.org/2024.eamt-1.32/

54. (Wolff, 2026) https://lokalise.com/blog/what-is-the-best-llm-for-translation/

55. (Khan, 2026) https://perplexityaimagazine.com/ai-tools/best-ai-for-translators-2026/

56. (Zhao et al., 2024) https://arxiv.org/html/2403.14118v2

57. (Nunziatini & Speroni, 2026) https://aclanthology.org/2026.eamt-2.24/

58. (UNESCO, 2025) https://www.unesco.org/en/articles/unesco-launches-global-roadmap-multilingualism-digital-era

59. (Adelani et al., 2024) https://aclanthology.org/2024.eacl-long.14/

60. (International Organization for Standardization, 2017) https://www.iso.org/standard/62970.html

61. (International Organization for Standardization, 2024) https://www.iso.org/standard/80701.html

62. (European Commission, 2026b) https://digital-strategy.ec.europa.eu/en/policies/guidelines-ai-transparency-obligations

63. (Sarti et al., 2022) https://aclanthology.org/2022.emnlp-main.532/

64. (Macken, 2024) https://aclanthology.org/2024.ctt-1.7/

65. (Briva-Iglesias & O'Brien, 2024) https://aclanthology.org/2024.eamt-1.37/

66. (Mihalache & Varela Salinas, 2026) https://aclanthology.org/2026.eamt-1.40/

67. (Velazquez et al., 2025) https://aclanthology.org/2025.acl-long.1569/

68. (Sarti et al., 2025) https://aclanthology.org/2025.tacl-1.64/

69. (Hu et al., 2025) https://aclanthology.org/2025.findings-emnlp.1289/

70. (Raunak et al., 2023) https://aclanthology.org/2023.findings-emnlp.804/

71. (Schneider et al., 2025) https://aclanthology.org/2025.findings-acl.500/

72. (Zhao et al., 2025) https://aclanthology.org/2025.findings-emnlp.1256/

73. (Guo et al., 2024) https://aclanthology.org/2024.lrec-main.1362/

74. (Lu et al., 2024) https://aclanthology.org/2024.emnlp-main.55/

75. (Compaore et al., 2026) https://aclanthology.org/2026.loreslm-1.53/

76. (Yankovskaya et al., 2023) https://aclanthology.org/2023.nodalida-1.77/

77. (Pashchenko et al., 2025) https://aclanthology.org/2025.nodalida-1.50/

78. (Chen et al., 2024) https://aclanthology.org/2024.eamt-1.20/

79. (Azime et al., 2026) https://aclanthology.org/2026.findings-acl.42/

80. (International Organization for Standardization, 2026) https://www.iso.org/standard/88184.html

81. (European Commission, n.d.-a) https://commission.europa.eu/resources-partners/etranslation_en

82. (European Commission, n.d.-b) https://commission.europa.eu/languages-our-websites/use-machine-translation-europa_en

83. (European Commission, 2026a) https://digital-strategy.ec.europa.eu/en/news/commission-publishes-code-practice-marking-and-labelling-ai-generated-content

84. (European Parliament and Council of the European Union, 2024)
https://eur-lex.europa.eu/eli/reg/2024/1689/oj