IN PRESS

MMSL
Military Medical Science Letters
Since 1925

Mil. Med. Sci. Lett. (Voj. Zdrav. Listy) 2025, 94, 1-7, On-line first
ISSN 2571-113X (Online)
DOI: 10.31482/mmsl.2025.008

ORIGINAL ARTICLE

EVALUATION FRAMEWORK FOR AI-BASED MACHINE TRANSLATION AND PROOFREADING TOOLS IN MEDICAL AND PHARMACEUTICAL WRITING

Josef Drahokoupil, Eva Drahokoupilová

Language Centre, University of Defence, Brno, Czech Republic

Received 9th July 2025.
Accepted 9th December 2025.
On-line 15th December 2025.

Summary

Recent advances in artificial intelligence (AI) have introduced powerful tools for machine translation (MT) and automated language proofreading (ALP) into academic publishing. However, their evaluation in highly specialized fields such as medicine and pharmacy remains methodologically underexplored. This study presents a multidimensional evaluation framework that integrates human expert judgment with AI-based replication to assess the quality of AI-generated translations and proofreading outputs. The framework covers three quality dimensions applicable to both MT and ALP: semantic fidelity, terminological accuracy and consistency, and grammatical correctness and fluency, as well as a fourth dimension specific to ALP, the appropriateness of edits.

In the pilot phase of the research, five machine translation systems were compared using the multidimensional evaluation framework, with DeepL serving as a strong baseline; under advanced idiomatic prompting, the Large Language Model (LLM) systems achieved performance levels comparable to this baseline. The semantic fidelity dimension was further evaluated through an AI simulation of human judgment using ChatGPT-5. The agreement between human and AI evaluators reached $\kappa = 0.81$ (95% CI = 0.73–0.88), indicating high consistency and no observed hallucinations within this pilot sample.

These preliminary results demonstrate that large language models have promising potential to reproduce human expert quality reasoning when guided by structured prompts. Beyond its technical contribution, our ongoing research represents a step toward transparent and reproducible evaluation for AI-generated academic writing and its automated quality assessment.

Key words: AI-assisted writing; machine translation; automated proofreading; evaluation Framework; semantic fidelity; terminological accuracy; COMET QE; QuickUMLS; biomedical language; scientific communication; domain-specific NLP; language quality assessment

Introduction

Over the past decade, artificial intelligence (AI) tools in machine translation (MT) and automated language proofreading (ALP), have become increasingly present in academic publishing, including the medical and pharmaceutical sciences. These fields, however, place exceptionally high demands on terminological accuracy, semantic

✉ University of Defence, Language Centre, Kounicova 65, 662 10 Brno, Czech Republic
josef.drahokoupil@unob.cz

IN PRESS

fidelity, and stylistic appropriateness (25,7). While AI tools significantly enhance productivity and accessibility, their use in highly specialized writing reveals a critical lack of validated methods and usage guidelines.

Despite their popularity, authors often use these tools without structured workflows, comparative evidence, or expert guidance (12,6). In our separate ongoing survey, 40 medical professionals have responded so far, with 58 % reporting the use of AI systems. About 65 % of these users rely on short, non-structured prompts and limited iteration, suggesting the absence of systematic prompt-design practices.

Benchmarking standards and best practices for academic contexts are also lacking (13,14), leaving users dependent on informal peer support, often at the cost of terminological or semantic precision (15,21). Our research has ambition to address this methodological gap. We propose a multidimensional evaluation framework that integrates structured expert assessment with methodological replication through large language models, while also providing a systematic approach to measuring MT and ALP quality and formulating evidence-based recommendations for academic writing.

Theoretical Background

Recent advances in natural language processing (NLP) have transformed the production of multilingual scientific texts. Tools such as DeepL, Google Translate, Grammarly, and ChatGPT are now widely used, yet their application to expert texts in medicine and pharmacy remains underexamined. Available studies tend to focus on translations of clinical reports or patient-facing materials (4), while scientific manuscripts targeting peer-reviewed journals receive less attention (6,21).

Proofreading tools are increasingly applied in academic writing, but are typically assessed based on surface corrections rather than their impact on meaning or style (1). Similarly, standard metrics such as BLEU or TER often misrepresent quality in specialized domains (8,10), while more sophisticated approaches like COMET QE and QuickUMLS (3,9) and use of LLM remain under-validated in scientific contexts (12).

Operational questions also remain unresolved, such as whether to translate text in full or sentence-by-sentence, how often to iterate corrections, and how to combine proofreading with translation (17,18). This forms the basis for a new generation of evaluation tools and academic best practices tailored to expert AI-assisted writing.

Selected Tools and Samples

To ensure representative and valid evaluations of language tool outputs, we selected tools reflecting diverse technological paradigms. For MT, we included DeepL and Google Translate (neural MT), ChatGPT, MS Copilot, Google Geminy (LLM) and ModernMT (adaptive MT) (20,5). For automated proofreading, we selected Grammarly (rule-based), LanguageTool (open-source), DeepL Write (LLM-powered), and ChatGPT, MS Copilot, Google Geminy (LLM) (1,2). We plan to explore the potential of translation aggregators such as MachineTranslation.com, which combine outputs from multiple engines. While these tools may produce higher-quality translations (14), their limited transparency and heterogeneous architecture raise methodological challenges.

For the purpose of ensuring statistically valid and comparable evaluations, 13 thematic-linguistic clusters were defined to represent the main biomedical and pharmaceutical domains (16,23). These clusters represent terminologically coherent domains constructed using a combination of semantic-similarity evidence, ontology-based proximity (UMLS/MeSH/SNOMED), shared pharmacotherapeutic and diagnostic vocabulary, and the typical sublanguage structures documented in biomedical discourse (16,23). To further ensure domain diversity and reproducibility, each cluster was cross-checked against the major Web of Science and Scopus subject areas relevant to that domain, which served as an external reference framework during article selection. A brief illustration is provided for one pilot cluster—cardiometabolic and vascular medicine—covers disciplines such as cardiology, diabetology, nephrology, and angiovascular medicine. These fields share overlapping diagnostic terminology (e.g., hemodynamic parameters), pharmacotherapy (e.g., RAAS-related agents, SGLT2 inhibitors), and conceptual structures. Its scientometric footprint spans several WoS and Scopus categories (Cardiac & Cardiovascular Systems; Endocrinology & Metabolism; Nephrology), ensuring that the sampled articles reflect genuine domain diversity.

IN PRESS

This stratified design provides statistically meaningful representation while maintaining cross-domain comparability and supports transferability of the method to future biomedical applications.

Articles will be sourced from Czech and international peer-reviewed journals. For the proofreading evaluation, synthetic texts will be generated in a controlled manner. Typical errors will be defined based on published analyses of scientific writing and existing error-correction corpora. Automated scripts will then introduce controlled linguistic deviations, such as article omission, verb agreement, or collocation errors, to simulate realistic author error patterns.

Sample size followed standard proportion-testing formulas for a 95 % confidence level and ±5 % margin of error, indicating that at least 385 sentence-level evaluation units are required to obtain statistically reliable inferences for a single evaluated outcome (11,22). As this article presents preliminary results, pilot analyses from the first cluster are used solely to verify the feasibility of the evaluation workflow, the internal logic of the method, and to provide initial evidence that LLMs can approximate human evaluators, which is essential for future automation. The full study will operate at a substantially larger scale, with approximately 56 full-length biomedical articles (≈4 per cluster), each assessed by two human experts and ChatGPT-5 across all tested Machine Translation and Proofreading tools. Statistical analysis in the full study will include ANOVA, the Friedman test, effect sizes, and inter-rater agreement measures (Cohen's κ) (18,17).

Evaluation Framework - Quality Assessment Method

The framework consists of four main quality dimensions, three of which are common to both translation and proofreading, and the fourth is specific to proofreading.

1. Terminological Accuracy and Consistency

This dimension evaluates whether domain-specific terms are translated or edited correctly according to established terminology, and whether they are used consistently throughout the text. In medicine and pharmacy, even minor terminological deviations can lead to misinterpretation. We assess both term accuracy and recurrence consistency.

2. Semantic Fidelity

This focuses on how well the output (translation or proofreading) preserves the original meaning. In translations, we assess cross-lingual meaning accuracy; in proofreading, we assess how faithful suggested edits are to the original message. It covers semantic shifts, omissions, and distortions.

3. Grammatical and Syntactic Correctness

This measures formal language quality—verb tense, agreement, word order, punctuation, and fluency. Highly technical texts must meet both scientific and linguistic standards. This dimension is critical to publishability.

4. Appropriateness of Suggested Edits (proofreading only)

In proofreading tasks, we assess whether the proposed changes are justified and beneficial. This dimension captures cases of "overediting"—where correct text is altered unnecessarily, potentially affecting authorial tone or message precision.

Each criterion is evaluated by two independent experts using an ordinal scale with commentary. In cases of disagreement, an adjudicator intervenes. Final scores are calculated as weighted averages, with separate default weight distributions for translation and proofreading tasks. For translations, the weights prioritize terminological accuracy and consistency (40 %), semantic fidelity (30 %), grammar correctness and language fluency (30 %). Domain-specific terminology represents the core of scientific validity and safety; even minor terminological shifts may alter meaning or clinical interpretation. For proofreading, the weights are assigned as follows: terminology (25 %), semantic fidelity (25 %), grammar and language fluency (30 %), and appropriateness of edits (20 %),

IN PRESS

the latter being specific to proofreading tasks. These weight parameters serve as configurable defaults, allowing users to adjust them according to their analytical priorities. To ensure such adaptability, we report all dimension-level results in this and future publications, enabling readers to recompute overall scores under alternative weighting schemes.

Preliminary results

The preliminary phase focused on the application of the proposed quality evaluation method to five MT systems. ChatGPT-5 and MS Copilot were tested using four prompt versions (Prompt 0–3). Prompt 0 represented a simple instruction ("translate"), which was subsequently refined into Prompt 3, based on our quality assessment framework integrating three quality dimensions. Special attention was paid to literal vs idiomatic translation.

For illustration:

- Literal translation: doporučená léčba → recommended medication
- Idiomatic (discipline-specific) translation: doporučená léčba → guideline-directed medical therapy, a domain-preferred term even though not a literal equivalent.

Our separate ongoing contextual survey among 43 physicians, conducted outside the scope of this study, indicates that a majority prefer idiomatic, discipline-specific translations, which supports the relevance of focusing on idiomatic renderings in our examples. The results achieved with our "advanced idiomatic" Prompt 3 demonstrate the potential to reach top-quality MT outputs.

The corpus used for testing consisted of a specialized medical article, "EMPEROR Reduced – Cardiac and Renal Outcomes with Empagliflozin in Patients with Heart Failure and Reduced Ejection Fraction," comprising 1,540 words and 65 sentences of domain-specific terminology (e.g., DM, HF, SGLT2, eGFR, NT-proBNP, ACEI, NYHA).

Human expert evaluation was complemented by an AI simulation of human assessment in the semantic fidelity dimension. Each sentence pair (source and translation) was evaluated by a human expert and by ChatGPT-5, which was prompted to emulate human judgment using the same 0–3 ordinal scale (0 = error-free, 1 = minor semantic inaccuracy, 2 = semantic divergence, 3 = semantic error). To ensure reproducible scoring, this scale is anchored by concise operational definitions and supported by a small set of illustrative examples that is gradually expanded through adjudication. Below we provide example of the boarders between level 1 = minor semantic inaccuracy and level 2 = semantic divergence:

- Level 1 denotes a small, non-critical meaning shift that preserves the clinical sense and would require only light post-editing.
- Level 2 indicates a meaning change that alters clinical interpretation, introduces a different concept, or modifies a key logical relation (actor, event, comparator).

The default prompt used in the preliminary results included only a minimal example set; this set is being progressively extended through adjudicated cases to form a growing database of human-validated decisions, enabling the LLM evaluator to converge toward stable, human-aligned judgments.

The results confirm that prompt design has a decisive impact on MT performance, particularly in semantic fidelity.

The advanced idiomatic prompt notably improved ChatGPT and MS Copilot outputs, achieving results comparable to DeepL. Google Gemini performed consistently well even with the basic prompt, demonstrating strong internal model optimization.

Human–AI agreement was quantified using linearly weighted Cohen's κ, as all observed discrepancies between raters were single-step differences. Because no two- or three-level divergences occurred, quadratic weighting would not yield additional insight and would artificially over-penalize disagreements absent in the dataset.

IN PRESS

Table 1. Indicative Comparison of Machine Translation Quality.

Machine Translation Tool	Grammar Human	Semantic Human	Semantic AI	Terminology Human	SCORE Translation Human
Deeple	85.25	81.97	81.97	98.21	89.45
Google Geminy (prompt - basic)	81.54	86.15	89.23	97.01	89.11
Google Geminy (prompt 3 - advanced idiomatic)	93.44	91.80	91.80	97.01	94.38
Google Translate	86.89	62.30	65.57	100.00	84.75
ChatGPT (prompt 0 - basic)	18.03	0	0	96.77	39.28
ChatGPT (prompt 3 - advanced idiomatic)	85.25	85.48	87.10	100.00	91.22
MS Copilot (prompt 0 - basic)	26.23	0	0	96.77	38.71
MS Copilot (prompt 3 - advanced idiomatic)	77.05	78.69	86.89	96.77	85.92

Across 560 sentence pairs, human–AI agreement on the 0–3 semantic fidelity scale reached $\kappa = 0.75$ (linearly weighted; 95 % CI = 0.68–0.82), indicating substantial to almost perfect alignment. The quadratically weighted coefficient was $\kappa = 0.82$, confirming robustness across weighting schemes. This high and statistically significant alignment shows that AI-based evaluation can effectively reproduce human expert judgment in semantic fidelity, confirming the feasibility of developing a fully automated system for machine translation quality assessment.

Discussion

The preliminary results confirm the methodological validity of the proposed evaluation framework and demonstrate its potential to align human and AI-based assessment. The high level of agreement ($\kappa = 0.81$) suggests that large language models can reliably approximate expert human judgment when guided by structured evaluation criteria.

During the development phase, we explored several automatic metrics, including COMET-QE, which theoretically measures sentence-level semantic adequacy. However, it proved impossible to establish threshold values that would allow consistent mapping of COMET scores to the human scale, and a weak negative correlation with human ratings ($r = –0.28$, $p = 0.028$) indicated that COMET-QE does not align closely enough with human ordinal scores to serve as a substitute for expert evaluation, and it was therefore excluded from the final workflow.

A notable observation emerged when comparing individual human and AI ratings. In several instances, the AI model correctly captured semantic nuances that the human evaluator initially overlooked. For example, in the sentence:

"Primární kombinovaný cíl byl kardiovaskulární úmrtí a hospitalizace pro zhoršení SS."
→ "The primary composite endpoint was cardiovascular death or hospitalization for worsening HF."

the human evaluator marked the translation as inaccurate due to the conjunction shift ("and" → "or"), whereas the AI correctly judged it as semantically accurate, reflecting standard clinical trial conventions. This interpretation is consistent with official EMPEROR-Reduced trial publications. Human annotation in this pilot phase was performed by a biomedical linguist with 43 years of professional experience in medical and pharmaceutical translation. This case illustrates how AI-based evaluation can complement expert human judgment by integrating domain conventions that may not always align with literal linguistic expectations.

What is particularly noteworthy is that the AI was able to articulate its reasoning clearly, justify its semantic choice, and when re-evaluated in subsequent iterations, maintain consistent logic without deviation or hallucination. This indicates that the model not only reproduces human-like judgments but also demonstrates a degree of interpretative stability and self-correction, which are essential for reliable automation of expert-level evaluation.

Such examples show that AI-based evaluators should demonstrate superior contextual consistency in domain-specific phrasing, especially when aligned with disciplinary conventions. In the next research phase, we plan to extend this approach to the remaining quality dimensions – terminology and grammar.

IN PRESS

The present results are based on pilot data, which limits the generalizability of the findings. At this stage, only the semantic fidelity dimension has been validated as a successful proof of concept for automated evaluation. The remaining dimensions – terminological accuracy and grammatical correctness – have not yet been tested for AI–human alignment and will require further validation on larger and more diverse corpora. Therefore, the reported agreement values should be interpreted as preliminary, indicating methodological potential rather than the framework's final performance across all evaluation dimensions.

Because no validated or consensus weighting scheme exists for MT or proofreading in biomedical writing, the default weights used in this study are best understood as arbitrary yet fully configurable parameters, serving as an initial benchmark rather than a fixed standard. The framework fundamentally relies on the individual quality dimensions, with the aggregated index used only for high-level orientation. Establishing evidence-based or domain-specific weighting schemes remains an open methodological question for future phases of the project.

The proposed framework follows a "human-in-the-loop" paradigm, in which AI acts as an augmentative evaluator rather than a replacement for human expertise. Such integration raises important ethical and procedural considerations regarding transparency, authorship, and accountability in evaluative workflows, as highlighted by Ben Saad *et al*. (2025), who describe this balance as the "assisted technology dilemma." (26).

Once a similarly high level of agreement between human and AI evaluators is achieved, we will move toward gradual automation of the assessment process—first by using AI as a second adjudicator to reduce turnaround time across all clusters, and subsequently by transitioning to full automation in the final stage. This step will enable the development of an applied research prototype – an application designed for automated translation quality evaluation based on validated human – AI agreement models.

Conclusion

Despite the rapid expansion of AI-assisted translation and proofreading tools, no validated methodology or benchmark study currently exists to guide medical and pharmaceutical professionals in selecting appropriate systems or using them effectively—including the design of prompts and workflows. This lack of methodological grounding leaves authors without reliable orientation regarding the quality, limitations, and best-use practices of MT and ALP systems.

This paper introduces a multidimensional framework that begins to fill this gap by providing a structured, reproducible method for evaluating AI-generated translations and proofreading outputs. Beyond the value of a single benchmark comparison, the key contribution lies in demonstrating that large language models can replicate expert semantic judgments with high reliability ($\kappa = 0.81$), thereby enabling the future automation of quality assessment.

Such automation has substantial practical implications. By reducing the time and expert effort required to verify translation and proofreading quality, the approach has the potential to accelerate the preparation of scientific manuscripts—an increasingly critical need in fast-moving clinical and research environments.

The next phase of the project will extend the evaluation to all biomedical clusters and all MT/ALP tools, refine weighting schemes, and support the development of an applied research prototype for automated assessment. The broader goal is to contribute to transparent, scalable, and domain-sensitive quality standards for AI-assisted academic writing.

Readers interested in obtaining the full wording of Prompt 3 (idiomatic) for LLM translation may access it through the QR code below after completing our ongoing survey, where participants can share their experiences, prompting practices, and workflow design, thereby contributing to the next phase of our research.

IN PRESS

References

1. Allen D, Mizumoto A. Automated grammar correction for academic writing: A critical review of tools and practices. J. Second Lang. Writ. 2024;63:101–119.
2. Barrot JS. Exploring the efficacy of automated writing feedback tools: A comparative study. Lang. Learn. Technol. 2022;26(1):45–62.
3. Batool S, Guo J, Müller T. Cognitive load in scientific writing: Measuring reader effort in machine-generated text. ACM Trans. Inf. Syst. 2024;42(2): Article 15.
4. Bui DD, Nguyen TN, Le TM. Evaluating neural machine translation of patient education materials. BMC Med. Inform. Decis. Mak. 2020;20(1):220.
5. Çetin O, Duran A. Evaluating adaptive machine translation systems: A comparative perspective. Mach. Transl. 2023;37(4):321–340.
6. Chowdhury S, Ghosh A, Sarkar D. Assessing AI-based proofreading tools in biomedical research writing. Comput. Biol. Med. 2022;142:105217.
7. Cillo R, Cortese A, Mariani J. Terminological precision in AI-assisted scientific translation: An Italian-English corpus study. Terminology. 2024;30(1):25–48.
8. Dahan L, Kalmanovich S, Goldberg Y. BLEU is not enough: Evaluating scientific translations beyond surface similarity. Trans. Assoc. Comput. Linguist. 2024;12:1–15.
9. Farber SM. Cognitive readability as a metric in technical writing: From perception to prediction. J. Tech. Writ. Commun. 2024;54(2):154–176.
10. Flores Marroquín CA. TER versus COMET: How automatic metrics fare in pharmaceutical translations. Rev. Lingüíst. Apl. 2024;60(2):113–130.
11. Fraser J, Heng MA. Measuring quality in medical translations: A term-based approach. Transl. Interp. 2014;6(1):23–38.
12. Huang Z, Tang C, Xu J. Validity and reliability of COMET-QE for technical texts. Comput. Linguist. 2025;51(1):87–109.
13. IDC. AI in academic language services: Trends and gaps. IDC Res. Rep. 2024; Q2.
14. Intento. State of machine translation: Industry benchmarks and trends. Intento Benchmark Rep. 2023.
15. Jaschke AC, Wang L, Niemeyer K. Machine-assisted writing in life sciences: Ethical and practical considerations. BioScience. 2024;74(1):33–40.
16. Kamdar MR, Musen MA, Shah NH. Clustering biomedical terminologies using UMLS: A graph-based approach. J. Biomed. Inform. 2017;68:31–46.
17. Knowles R, Lo CK. Multidimensional quality metrics in low-resource scientific domains. Proc. EACL. 2024:211–221.
18. Licht R, Becker S, Kohl M. Evaluating annotation consistency in translation quality assessment. Lang. Resour. Eval. 2022;56(2):789–804.
19. Liu M, Liu J, Yu H. Topic modeling for clustering medical specialties in large-scale literature. J. Am. Med. Inform. Assoc. 2012;19(2):220–226.
20. López Caro A. ModernMT: A neural adaptive system for domain-specific translation. In: Proc. Mach. Transl. Summit. 2023:41–50.
21. Muhanna M. Human-AI collaboration in academic writing: Evidence from pharmacy journals. J. Scholarly Publ. 2025;56(3):222–241.
22. Shah A, Cunningham J, Murchison T. Scientific sentence segmentation and its impact on translation accuracy. Lang. Resour. Eval. 2016;50(4):933–952.
23. Soares F, Ribeiro R, Wanner L. Domain clustering using word embeddings: Application to biomedical literature. Nat. Lang. Eng. 2019;25(3):325–343.
24. Translated IDC. Adaptive MT and professional workflows: Report on use and perception. IDC Tech. Brief Ser. 2022.
25. Wong KY, Symons T. Terminology control in AI-powered academic editing tools. Terminol. Knowl. Eng. 2025;29(1):87–103.
26. Ben Saad H, Dergaa I, Ghouili H, et al. The assisted technology dilemma: a reflection on AI chatbots use and risks while reshaping the peer review process in scientific research. AI & Society. 2025;1:1–8.