ISSN 2669-2023 (print) / ISSN 2538-8959 (on-line) / **https://doi.org/10.47459/cndcgs.2026.71**

A Multi-Criteria LLM-Based Framework for Automated Evaluation of Scientific Language Quality

Josef DRAHOKOUPIL[1], Eva DRAHOKOUPILOVÁ[1] , Jana VAŇKOVÁ[1], Dana FIALOVÁ[1]

[1]*Language Centre, University of Defence, Kounicova 65, 662 10, Brno Czech Republic*

*Correspondence:*eva.drahokoupilova@unob.cz*

Abstract

This study proposes a multi-criteria framework for automated evaluation of scientific language quality using Large Language Models (LLMs). The approach models translation assessment as a structured decision process across semantic, grammatical, and terminological dimensions aligned with ISO 5060:2024. Validation on biomedical data shows strong agreement with expert judgment ($\kappa = 0.81$), while metric-based estimation demonstrates weak correlation. The results indicate that interpretable LLM-based evaluation can reliably replicate expert reasoning and support scalable, transparent language quality assessment.

KEY WORDS: *multi-criteria evaluation; structured decision architecture; Large Language Models; editorial proofreading services; inter-rater agreement; prevalence-adjusted kappa; scientific language assessment; AI calibration*

Citation: Drahokoupil, J.; Drahokoupilová, E.; Vaňková, J.; Fialová, D. A Multi-Criteria LLM-Based Framework for Automated Evaluation of Scientific Language Quality. In Proceedings of the Challenges to National Defence in Contemporary Geopolitical Situation, Brno, Czech Republic, 7-10 September 2026. ISSN 2538-8959, https://doi.org/10.47459/cndcgs.2026.71

1. Introduction

The evaluation of scientific language quality represents a complex multi-criteria decision-making (MCDM) problem. In specialized domains such as medicine and pharmaceutical research, translation and proofreading quality cannot be reduced to a single numerical metric, as multiple interdependent dimensions—semantic fidelity, terminological accuracy, grammatical correctness, and domain-specific idiomaticity—must be assessed simultaneously. These criteria often interact, requiring structured prioritization and consistent interpretative logic.

While artificial intelligence (AI) tools for machine translation (MT) and automated language proofreading (ALP) are increasingly integrated into academic workflows, current evaluation approaches rely predominantly on automatic metrics (e.g., COMET-QE [11], BLEURT [9]) that provide opaque scalar outputs without transparent diagnostic interpretation. Such metrics may correlate with human judgment at the corpus level but fail to support structured, criterion-based decision processes necessary in high-stakes scientific communication.

This study proposes a multi-criteria evaluation framework operationalized through Large Language Models (LLMs), designed to replicate expert-level reasoning within an explicitly defined rating matrix aligned with ISO 5060:2024 principles. By embedding structured evaluation descriptors into LLM prompts, the framework transforms language-quality assessment from heuristic scoring into a reproducible decision architecture capable of scalable and interpretable automation.

The primary objective of this research is to demonstrate that LLM-guided evaluation can function as a transparent multi-criteria decision-support system, achieving reliable alignment with expert human judgment while preserving diagnostic clarity across linguistic dimensions.

2. Method of investigation

The study conceptualizes translation quality assessment as a structured multi-criteria decision-making (MCDM) process grounded in established translation-evaluation theory and recent advances in LLM-based assessment [14,15]. The present contribution focuses on the translation evaluation model; a related proofreading evaluation model has been developed within the broader research program but is not discussed here in detail.

Peer-review under responsibility of General Jonas Žemaitis Miltitary Academy of Lithuania and University of Defence, Czech Republic

The framework is conceptually aligned with ISO 5060:2024 principles for structured evaluation of translation output [16] and reflects the logic of multi-dimensional quality models such as MQM. Each dimension operates on an ordinal decision scale; however, the qualitative interpretation of levels is defined separately for every criterion.

The proposed approach decomposes translation quality into three core dimensions: semantic fidelity, grammatical correctness and fluency, and terminological precision and consistency. These dimensions are evaluated independently and subsequently integrated into an overall quality score.

The evaluation process begins with segmentation of the source and translated texts into aligned sentence units, followed by identification and mapping of domain-specific terminology. This structured alignment enables parallel assessment at both segment level and term level.

The overall architecture of the proposed evaluation framework is illustrated in Figure 1, providing a conceptual overview of the evaluation workflow and the interaction between individual quality dimensions.

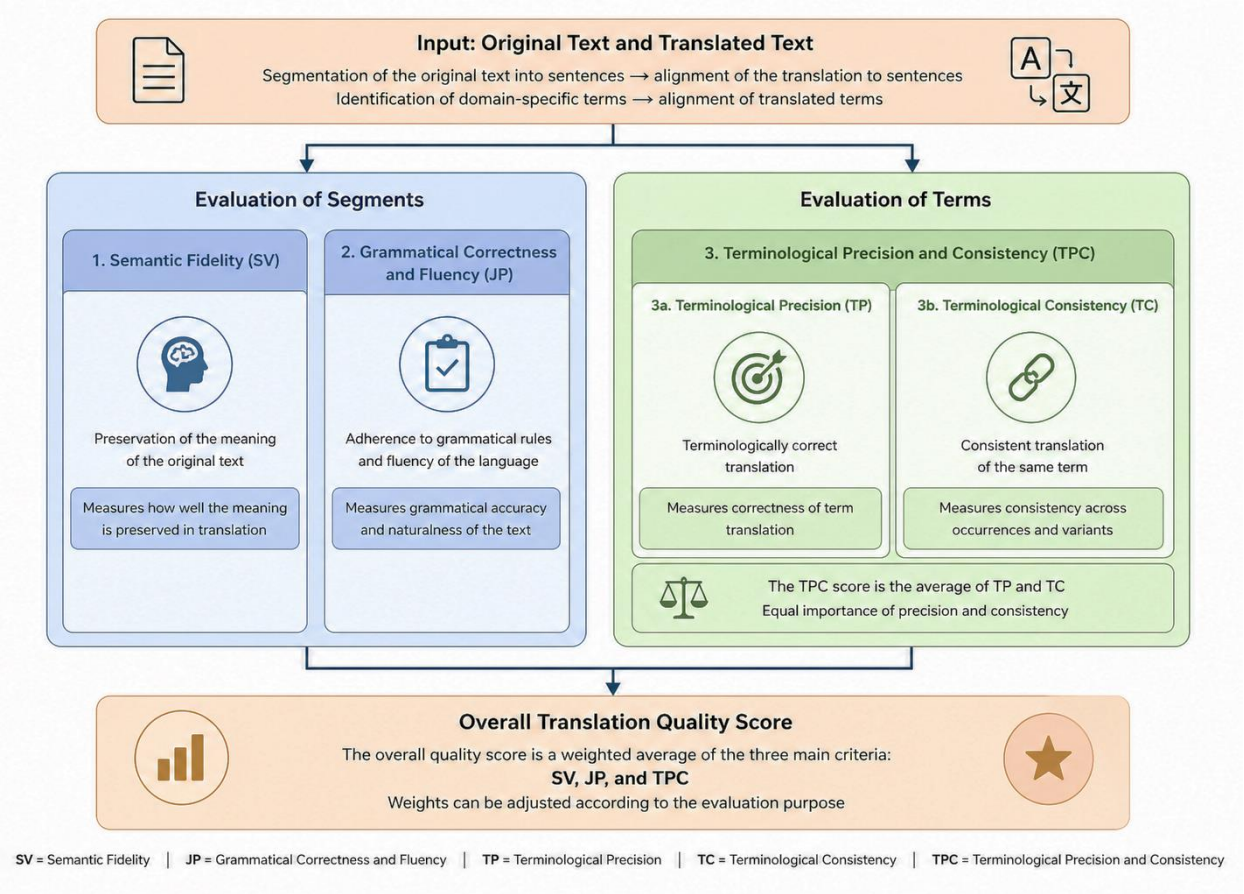

Fig. 1. Translation quality framework.

2.1 Sub-metric Computation

Each evaluation dimension is computed based on observable error patterns identified during the assessment process. To ensure interpretability and comparability, all sub-metrics are normalized to a 0–100 scale, where higher values indicate higher translation quality.

Semantic Fidelity (SV). Semantic fidelity reflects the degree to which the meaning of the source text is preserved in translation. It is computed as:

$$SV = 100 - \left(\frac{1 \cdot N_1 + 2 \cdot N_2 + 3 \cdot N_3}{N_{seg}} \cdot 100\right) \quad (1)$$

where N_1represents minor semantic inaccuracies, N_2semantic deviations, N_3semantic errors, and $N_{seg}$the total number of evaluated segments.

Grammatical Correctness and Fluency (JP). Grammatical correctness and fluency capture adherence to grammatical rules and overall linguistic well-formedness of the translated text. The metric is computed analogously:

$$JP = 100 - \left(\frac{1 \cdot N_1 + 2 \cdot N_2 + 3 \cdot N_3}{N_{seg}} \cdot 100\right) \quad (2)$$

where N_1, N_2, and N_3denote minor, major, and severe grammatical errors, respectively.

Terminological Precision (TP). Terminological precision measures the correctness of individual term translations:

$$TP = 100 - \left(\frac{1 \cdot N_1 + 2 \cdot N_2}{N_{term}} \cdot 100\right) \quad (3)$$

where N_1represents minor terminological deviations, N_2incorrect term translations, and $N_{term}$the number of evaluated terms.

Terminological Consistency (TC). Terminological consistency evaluates whether identical source terms are translated consistently across the text. It is computed as:

$$TC = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\max(f_{variant})}{\Sigma f_{variant}} \cdot 100\right) \quad (4)$$

where $f_{variant}$denotes the frequency of individual translation variants and $K$the number of evaluated terms.

Terminological Precision and Consistency (TPC). The overall terminological score combines precision and consistency:

$$TPC = \frac{TP + TC}{2} \quad (5)$$

The 0–100 scale is used for aggregated metric computation, while the 0–3 scale is applied at the segment level for interpretative evaluation.

Criterion-specific descriptors were embedded into a structured LLM evaluation prompt. This approach follows recent findings Large Language Models can reproduce human-like evaluation logic when provided with explicit decision boundaries [14,15]. Unlike metric-based quality estimation systems such as COMET-QE or BLEURT [11–13], the proposed framework does not rely on scalar similarity scores but on transparent, rule-guided reasoning.

2.2 Framework validation

The following section presents the empirical validation of the proposed framework. Inter-Rater Reliability Validation - To examine whether LLM-guided evaluation can replicate expert judgment, a dedicated validation subset of 245 Czech-English biomedical sentence pairs was used exclusively for inter-rater agreement analysis. Each sentence pair was independently evaluated by:

• a human domain expert,

• a Large Language Model instructed to apply identical criterion-specific descriptors.

Agreement in the semantic dimension was assessed using weighted Cohen's kappa (κ) [6] with linear weights, appropriate for ordinal multi-criteria data. This procedure enables robust measurement of alignment between human and AI evaluative reasoning.

Comparative Metric Analysis - To assess whether metric-based estimation reflects structured expert evaluation, a subset of sentence pairs was additionally analyzed using COMET-QE [8,11,13]. Pearson correlation analysis was performed between human ordinal ratings and COMET scalar outputs. This comparison allows examination of the relationship between interpretable multi-criteria evaluation and numerical quality-estimation metrics.

Empirical Needs Assessment - To establish the practical relevance of the proposed evaluation framework, an empirical survey was conducted among academic and research staff at Czech medical and pharmaceutical faculties. The survey examined:

language-production workflows (direct English writing vs. Czech-first then translation),

• use of AI-based and non-AI language tools,

• reliance on professional proofreading services,

• journal requirements regarding native-speaker certification.

The purpose of this survey was not to evaluate translation quality directly, but to determine whether a systematic, scalable, and interpretable evaluation framework is practically justified within contemporary academic publishing workflows.

3. Investigation Results

The empirical survey conducted among academic and research staff (n = 87) confirmed that language preparation remains a structurally significant component of scientific publishing. While 74% of respondents draft manuscripts directly in English, 26% primarily translate from Czech.

In a follow-up survey (n = 40), 60% reported that their target journals require formal confirmation of native-speaker proofreading prior to acceptance. Reported costs most frequently ranged between 5,000–7,999 CZK, with turnaround times

typically between one and four weeks. A complementary review of 60 high-impact biomedical journals showed that approximately 90% either require or strongly recommend native-speaker review [1–3], and around 40% formally recognize certified editing services [4,5].

These findings indicate that editorial workflows remain dependent on externally validated human language correction processes that are both time-consuming and financially demanding, thereby constituting a structural bottleneck in contemporary academic communication.

Inter-Rater Reliability of LLM-Based Evaluation - Agreement between human expert judgment and LLM-guided evaluation was assessed using linearly weighted Cohen's kappa (κ). For the semantic dimension (245 sentence pairs), the analysis yielded: $\kappa = 0.81$ (95% CI: 0.73–0.88; $p < 0.05$). This indicates strong alignment between structured LLM-based evaluation and expert reasoning. According to commonly used interpretation thresholds [7], this level corresponds to substantial to almost perfect agreement. To further illustrate the behaviour of the proposed evaluation framework, representative examples of LLM-based semantic assessment are provided in Table 1. The examples demonstrate how the model applies the ordinal evaluation scale (0–3) and generates concise, criterion-aligned justifications for its decisions.

Table. 1. LLM Justification capability.

Source text	MT Translation	LLM Evaluation	LLM Justification
Studie byla definována, že bude ukončena po dosažení 841 primárních cílů.	Google Translation: The study was defined to be terminated after reaching 841 primary endpoints.	2 – Semantic deviation	The translation is understandable but not idiomatic and partially inaccurate in meaning. The phrase "was defined to be terminated" is not standard in scientific English and sounds unnatural when describing protocol-defined study completion. The appropriate expression should be "was designed to be concluded" or "was planned to end", which accurately reflects how the study's termination point was specified in the protocol.
Studie byla definována, že bude ukončena po dosažení 841 primárních cílů.	DeepL Translation: The study was defined as ending after reaching 841 primary endpoints.	1 – Minor semantic inaccuracy	The translation conveys the general meaning, but "was defined as ending" is not idiomatic in scientific English. The phrase sounds awkward and uncommon in research reporting. A more natural rendering would be "The study was designed to end after 841 primary endpoints" or "The study was planned to conclude after 841 primary endpoint events."

For the grammatical dimension, evaluated on an initial dataset of 187 sentence pairs, linearly weighted κ reached: $\kappa = 0.60$ (95% CI: 0.39–0.75). Observed agreement was 90.3%. Given the strong prevalence of the "0" category, the prevalence-adjusted bias-adjusted kappa (PABAK) was additionally calculated:

PABAK = 0.81. The discrepancy between κ and PABAK reflects class imbalance rather than structural inconsistency of the model.

Human–Human Variability and the Role of Adjudication - A parallel grammatical evaluation performed by two independent human evaluators (218 valid sentence pairs), without adjudication, produced high observed agreement (81.6%) but very low κ values due to extreme prevalence imbalance.

This finding demonstrates that divergence in fine-grained grammatical judgment occurs even among experts applying identical methodological descriptors. In standard annotation practice, such discrepancies are resolved through adjudication procedures aimed at consensus formation.

Within the proposed framework, iterative prompt refinement serves as an analogue to adjudication. Disagreement patterns between evaluators and the LLM inform controlled adjustment of decision descriptors, progressively increasing alignment and stabilizing evaluation boundaries. Rather than indicating methodological weakness, preliminary disagreement functions as a diagnostic calibration phase within the multi-criteria architecture.

Comparison with Metric-Based Estimation - Human ordinal ratings were compared with COMET-QE scores [12,13]. Correlation analysis revealed: $r = -0.28$ ($p = 0.028$). The weak negative correlation indicates that scalar metric-based estimation does not reliably capture structured, criterion-specific decision logic.

Taken together, the results demonstrate that while editorial workflows remain dependent on slow human validation processes, structured LLM-based multi-criteria evaluation should replicate expert semantic reasoning with high reliability and can be systematically calibrated in dimensions exhibiting greater interpretative variability.

4. Conclusions

In an era of rapidly accelerating AI-supported scientific production, editorial review processes remain comparatively slow, structurally dependent on external native-speaker validation, and frequently associated with significant financial and temporal costs. As demonstrated by the empirical survey, such requirements continue to extend publication timelines and create systemic bottlenecks in highly specialized academic communication.

The present study contributes to addressing this imbalance by proposing a structured multi-criteria framework for evaluating the quality of highly specialized scientific texts. Rather than relying on opaque scalar metrics, the method formalizes linguistic assessment as an interpretable decision architecture grounded in explicitly defined evaluation dimensions.

Preliminary validation results indicate promising alignment between human expert judgment and LLM-guided evaluation, particularly in the semantic dimension. Variability observed in grammatical assessment highlights the inherently interpretative nature of linguistic review and underscores the importance of iterative calibration. The incorporation of adjudication-informed prompt refinement represents an initial step toward stabilizing decision boundaries and progressively improving evaluative consistency.

While further dataset expansion and statistical validation are necessary, the current findings suggest that structured AI-assisted multi-criteria evaluation may offer a viable pathway toward accelerating pre-submission quality assessment without compromising transparency or methodological rigor.

Once satisfactory calibration stability is achieved, the framework may provide the foundation for scalable and potentially commercializable language quality evaluation services, thereby reducing costs and shortening review cycles for highly specialized scientific texts.

5. Limitations

The present findings are based on preliminary datasets of limited size, and full statistical power has not yet been achieved across all evaluated dimensions. The validation has so far been conducted within a single biomedical linguistic cluster, which may limit generalizability to other domains. Additionally, not all evaluator discrepancies have undergone formal adjudication at this stage, and iterative calibration remains ongoing. Further dataset expansion and cross-domain validation will be necessary to confirm the robustness and transferability of the proposed framework.

References

1. Elsevier. Guide for Authors. Available from: https://www.elsevier.com/authors
2. Springer Nature. English Language Editing Services. Available from: https://www.springernature.com/gp/authors/research-data-policy
3. Wiley. Manuscript Preparation Guidelines. Available from: https://authorservices.wiley.com
4. American Medical Writers Association (AMWA). English Editing and Author Services. Available from: https://www.amwa.org
5. European Association of Science Editors (EASE). Guidelines for Authors and Translators. Available from: https://ease.org.uk
6. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174.
8. Rei R, Stewart C, Farinha AC, Lavie A. COMET: A Neural Framework for MT Evaluation. Proc EMNLP. 2020:2685–2702.
9. Sellam T, Das D, Parikh A. BLEURT: Learning Robust Metrics for Text Generation. Proc ACL. 2020:7881–7892.
10. Freitag M, Mathur N, Brown C, et al. Results of the WMT21 Metrics Shared Task. Proc WMT. 2021:622–640.
11. Rei R, Stewart C, Farinha AC, Lavie A. COMET-QE: Estimating MT Quality without References. Proc WMT. 2021:539–550.
12. Fomicheva M, Specia L. Reference-free MT evaluation: A systematic study. Compute Speech Lang. 2022;72:101302.
13. Kocmi T, Federmann C. Large Language Models are State-of-the-Art Evaluators of Translation Quality. Proc WMT. 2023.
14. Drahokoupil J, Drahokoupilová E. Evaluation Framework for AI-Based Machine Translation and Proofreading Tools in Medical and Pharmaceutical Writing. Mil Med Sci Lett. 2025;94:1–7. doi:10.31482/mmsl.2025.008
15. OpenAI. GPT-4 Technical Report. 2023. Available from: https://arxiv.org/abs/2303.08774
16. International Organization for Standardization. ISO 5060:2024 Translation services — Evaluation of translation output. Geneva: ISO; 2024.