

Evaluating Methods for Assessing LLM-Translated Clinical Discharge Instructions: A Comparison of Automated Metrics, MQM-Based Evaluation, and Clinician Review

William Mundo

william.mundo@cuanschutz.edu

University of Colorado School of Medicine

Shiyue Hu

University of Colorado School of Medicine

Alexander L Lupi

Denver Health and Hospital Authority

Elizabeth Goldberg

University of Colorado School of Medicine

Yanjun Gao

University of Colorado School of Medicine

Article

Keywords: Large-language models, Spanish discharge instructions, clinical translation evaluation, multidimensional quality metric

Posted Date: August 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10631294/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Large language models (LLMs) have emerged as promising tools for translating patient-facing clinical communication, yet robust methods for evaluating translation quality remain lacking. Existing evaluation approaches, including automated metrics, structured quality frameworks, and clinician review, capture different aspects of translation quality, but their agreement and complementary roles remain poorly understood.

Objective

Our objective was to compare automated metrics, MQM-based LLM-as-a-judge evaluation, and bilingual clinician review to 1) assess the quality of LLM-generated Spanish discharge instructions and 2) to identify domains in which human evaluation remains necessary.

Methods

We conducted an evaluation using 200 English-language hospital discharge instructions randomly selected from the MIMIC database. Three LLMs (Azure OpenAI GPT-5, GPT-4.1, and Gemma-4-12B-it) generated Spanish translations under clinician-framed and literal prompting conditions. Translation quality was assessed using automated fidelity metrics (BLEU, ROUGE-L, NIST, word error rate, and concept-based similarity), a multi-agent MQM-based LLM-as-a-judge framework, and blinded bilingual clinician review of a randomly selected subset. Clinician assessment served as the reference standard for agreement analyses using Cohen's kappa.

Results

Automated metrics differentiated translation performance across models and prompting strategies, while MQM-based evaluation demonstrated greater sensitivity to domain-specific differences in accuracy, terminology, fluency, and style than to formatting or localization. Agreement between LLM-based MQM scoring and clinician assessment was generally slight across domains ($\kappa = -0.18$ to 0.20), whereas inter-rater agreement between clinicians ranged from slight to moderate ($\kappa = -0.11$ to 0.49), highlighting the inherent subjectivity of clinical translation assessment. Despite high overall translation quality, we observed the greatest variability in domains requiring interpretation of clinical meaning.

Conclusions

LLM-generated Spanish discharge instructions demonstrated variable translation quality depending on the evaluation approach, with accuracy and terminology remaining important areas for clinician assessment. Automated fidelity metrics, MQM-based evaluation, and clinician review captured complementary rather than interchangeable dimensions of translation quality. A layered evaluation strategy combining scalable automated and structured assessment with targeted human review may provide a more comprehensive framework for evaluating patient-facing clinical translations prior to clinical deployment.

INTRODUCTION

Written clinical communication that is translated from English to Spanish plays a central role in patient safety, documentation integrity, and post-discharge care. Accurate translations can facilitate transitions of care, as language discordance has been linked to lower patient comprehension, higher rates of adverse events, and disparities in follow-up and medication adherence, especially among patients with limited English proficiency (LEP).^{1–3} Translated discharge instructions, in this context, represent a high-stakes form of communication in which translation quality can directly shape patient understanding and subsequent health behaviors.⁴

With the growing integration of large language models (LLMs) into clinical workflows, interest in their use for multilingual clinical communication has expanded. Early evaluations suggest that LLM-generated translations of discharge instructions can achieve high levels of sentence-level accuracy.^{5,6} At the same time, these systems are not uniformly reliable. A subset of translations may still contain errors that may carry clinical significance, particularly when nuances in meaning, terminology, or context are not preserved. Even infrequent errors in this setting may be consequential, which highlights the need for evaluation approaches that reflect clinical use rather than surface accuracy of translation alone.

A persistent challenge in this space is assessing translation quality. Automated metrics are widely used because they are scalable and reproducible, but they primarily reflect lexical overlap and structural similarity. Prior work has shown that these metrics do not consistently align with human judgment and may fail to capture errors that affect clinical meaning or interpretation.^{7,8} As a result, reliance on automated scoring alone may provide an incomplete picture of translation quality for patient-facing materials.

Frameworks such as the Multidimensional Quality Metric (MQM) offer a more structured approach by evaluating translation across multiple domains and categorizing errors by type and severity.⁹ In practice, however, the use of MQM in healthcare settings has been variable, with differences in scoring approaches, rater expertise, and reporting of agreement that make comparisons across studies difficult.^{10,11} More recently, LLM-based “judge” frameworks have been proposed to scale structured evaluation, raising the possibility of combining domain-based assessment with automation.¹² Whether these approaches are comparable to bilingual human judgement remains uncertain.

In this study, we evaluated the Spanish translation of discharge instructions using automated fidelity metrics, MQM-based LLM-as-a-judge evaluation, and bilingual clinician review. Our objective was to compare these evaluation approaches, examine their agreement with clinician assessment, and identify the strengths and limitations of each approach for evaluating patient-facing clinical translations.

METHODS

Study Design

We conducted a comparative evaluation of methods for assessing the quality of Spanish translations of discharge instructions. The study examined three complementary evaluation pathways: automated translational fidelity metrics, structured MQM-based LLM-as-judge scoring, and clinician review (Fig. 1). This study used de-identified, publicly available clinical text and involved no direct interaction with human subjects. It was determined to be non-human subject research (COMIRB 24-2170).

Data Source and Sample Selection

English-language discharge instructions were drawn from the Medical Information Mart for Intensive Care (MIMIC) database using a staged sampling strategy that balanced representativeness, content validity, and feasibility. We randomly sampled a pool of 200 discharge instructions to capture heterogeneity in clinical topic, length, and linguistic complexity. Because the primary objective of this study was methodological evaluation rather than hypothesis testing, no formal sample size calculation was performed. Instead, the corpus size was selected a priori to provide a diverse and representative set of patient-facing discharge instructions while maintaining feasibility for translation generation, structured MQM evaluation, and clinician review. This approach enabled evaluation across a broad range of instructional types and clinical content, supporting assessment of domain-specific error patterns and agreement among evaluation methods.

Translation Generation

Each discharge instruction was translated from English to Spanish using an LLM-based translation system using the prompt shown in Appendix A (**Figure S1**). We evaluated three models: GPT-4.1, GPT-5 and gemma-4-12B-it. Experiments were conducted on the University of Colorado Anschutz Azure environment (GPT-5, GPT-4.1) and a HIPAA-compliant, on-premises server with two H100 96GB GPUs (supporting Gemma-4-12B-it) managed by Office of Information and Technology, ensuring the data privacy and confidentiality. We generated translated outputs without post-editing. We did not include professionally translated human-written discharge instructions as a comparator. Consequently, this study evaluates agreement among translation evaluation methods rather than comparative performance against a professional translation benchmark.

To examine how prompt design influences translation quality, each model was evaluated under two prompting conditions: a clinically grounded prompt and a minimal literal prompt. The clinically grounded prompt (referred to as the “basic prompt”) framed the model in an expert role: “You are a clinician in the ICU responsible for translating English-language discharge instructions into Spanish,” followed by the source text and the instruction “As the clinician, translate the given Discharge Instructions into Spanish.” The minimal prompt (referred to as “the minimal prompt”) provided no role framing and simply instructed the model to perform a mechanical substitution: “Translate the following text word-for-word into Spanish.” This contrast was designed to isolate the effect of clinical context and role assignment on translation fidelity, fluency, and terminology accuracy. For each condition, a back-translation step was applied uniformly using the same clinician role prompt, converting the Spanish output back into English so that automated reference-based metrics could be computed against the original source text. The back-translation prompt itself was kept consistent across both conditions, as it serves as a measurement instrument rather than an experimental variable.

Automated Translational Metrics and Evaluation

Automated evaluation was performed using established translational fidelity metrics applied to English back-translations of the Spanish outputs. Back-translation refers to translating the Spanish output back into English to assess preservation of meaning relative to the original English discharge instruction. Automated metrics included Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and National Institute of Standards and Technology (NIST) scores, which quantify n-gram overlap between generated and ground-truth texts.^{13–15} We also computed the word error rate (WER) to quantify insertions, deletions, and substitutions in the back-translated text.¹⁶ These metrics have served as standard measures in natural language processing for decades, but were originally developed for general-domain tasks such as machine translation (BLEU, NIST), summarization (ROUGE), and speech recognition (WER) rather than for clinical text. To assess semantic alignment beyond surface-level overlap, we included a concept-based similarity (CUI) metric that compares extracted clinical concepts across back-translated outputs and the original English text, using standardized procedures and medical vocabulary from National Library of Medicine’s UMLS for concept extraction and matching.^{17,18} All automated metrics were computed using reproducible, publicly available implementations and summarized descriptively for comparative analysis across models and translation sources. COMET was not used because it is a source-aware neural metric intended for cross-lingual (source-to-target) evaluation, whereas our back-translation approach produces a monolingual English comparison; COMET also lacks validation in the clinical domain and yields an opaque score without the concept-level interpretability afforded by our overlap- and CUI-based metrics.

MQM Evaluation

MQM evaluation focused on domains pertinent to patient-facing discharge instructions, including accuracy, terminology, linguistic convention, style, local conventions, audience appropriateness, and design & markup (Table 1). MQM domains were operationalized using an ordinal, five-level severity scale: 5 = no error; 4 = minor; 3 = moderate; 2 = major; 1 = critical (meaning-altering or plausibly harmful). Full MQM scoring criteria and reviewer instructions are provided in Appendix B (Table 1S). MQM scoring was performed using an LLM-as-a-judge approach, in which an LLM was prompted with the MQM rubric as shown in Appendix C (**Figure S2**). The LLM produced domain-level severity ratings for each translation. Rubric definitions and scoring instructions were held constant across evaluations to ensure consistency. To establish clinical validity and serve as a benchmark for LLM-based evaluation, a bilingual clinician independently reviewed translated discharge instructions.

Table 1
Multidimensional Quality Metric Domain Operational Definitions

Domain	Operational Definition
Accuracy	Preservation of clinical meaning between source and translated text.
Terminology	Appropriate translation of medical terminology, medications, procedures, and diagnoses.
Linguistic convention	Correct grammar, spelling, syntax, punctuation, and sentence construction.
Style	Preservation of tone, formality, and communicative intent.
Local convention	Adherence to conventions expected in Spanish-language healthcare communication.
Audience appropriateness	Suitability for the intended patient audience.
Design & markup	Preservation of structure, formatting, lists, section headers, and organizational elements.
<i>Definitions reflect the operationalization of MQM domains used in this study.</i>	

To evaluate translation quality, we developed a multi-agent scoring pipeline grounded in the MQM framework. Rather than relying on a single model pass, we decomposed the evaluation into three sequential agents, each with a distinct role. The first agent performs error detection, reading the source English text alongside the Spanish translation and identifying candidate errors across seven MQM domains: accuracy, fluency, style, terminology, locale convention, audience appropriateness, and design markup. Each error is assigned a preliminary severity of critical, major, or minor, with the agent explicitly instructed not to penalize translations for grammatical irregularities or clinical shorthand already present in the source. The second agent acts as a verification layer, reviewing the candidate errors and filtering out false positives using both bilingual clinical reasoning and UMLS concept identifiers (CUIs) extracted via QuickUMLS.¹⁹ CUI evidence was used to ground the source span in a known biomedical concept, after which the agent applied its own clinical knowledge to judge whether the Spanish rendering

correctly expressed that concept, a deliberate design choice to prevent over-reliance on lexical matching. Finally, the scoring agent synthesizes the verified error list into a 1 to 5 score for each MQM domain, with explicit constraints ensuring that any domain with confirmed errors cannot receive a perfect score. All three agents were backed by OpenAI Azure GPT-5.4, called asynchronously via a LangGraph state graph to maintain structured state across the pipeline.

Clinician Review

To provide a clinically grounded reference standard, bilingual clinician MQM review was used as the benchmark against which automated and LLM-based evaluation approaches were compared. Clinician review was selected as the reference standard because the primary outcome of interest was translation quality in patient-facing communication, which ultimately depends on human interpretation of clinical meaning, linguistic appropriateness, and usability. Two human reviewers who practice as MDs with native Spanish-speaking and writing proficiency and bilingual provider certifications were blinded to automated metric scores and LLM-as-judge outputs and performed MQM scoring on a random subset of 20 translations to assess concordance with LLM-as-judge ratings and to evaluate feasibility before scaling to the full 200-dataset.

Agreement Analysis

Agreement analyses were performed to assess the extent to which MQM-based LLM-as-a-judge scoring approximated clinician MQM assessment across domains. For MQM domains, agreement was quantified using Cohen's kappa. Domain-specific agreement was examined to identify areas of convergence and divergence, particularly in domains requiring interpretation of clinical meaning.

Data Analysis

Analyses were prespecified to characterize agreement, discrimination, and domain-level performance across evaluation approaches rather than to test a single null hypothesis. Continuous measures are reported as means with standard deviations or medians with interquartile ranges, as appropriate, and categorical variables as counts and percentages. All analyses accounted for the paired structure of the data, with each discharge instruction contributing corresponding translated outputs across evaluation pathways. Pairwise comparisons between prompt conditions (clinician-framed versus literal) were conducted using the Wilcoxon signed-rank test applied to per-instance scores. Automated surface-level and concept-based metrics (BLEU, WER, NIST, ROUGE-L, and CUI F1) were computed using the NLTK, jiwer, rouge, and QuickUMLS Python libraries. MQM-based quality scores were derived from a multi-agent LLM-as-a-judge pipeline implemented in Python using LangGraph. All analyses were performed in Python 3.11.

RESULTS

Automated Translation Evaluation

Across all automated translational fidelity metrics, prompting strategy produced larger differences than model selection (Table 2). For each of the three evaluated models (GPT-5, GPT-4.1, and Gemma-4-12B-it), the literal prompt achieved higher BLEU, NIST, ROUGE-L, and CUI F1 scores and lower word error rates (WER) than the clinician-framed prompt (all Wilcoxon $p < 0.001$). For GPT-5, BLEU increased from 0.27 ± 0.11 to 0.38 ± 0.13 , NIST from 4.01 ± 1.04 to 4.90 ± 1.13 , ROUGE-L from 0.61 ± 0.09 to 0.70 ± 0.09 , and CUI F1 from 0.77 ± 0.07 to 0.82 ± 0.06 , while WER decreased from 0.73 ± 0.14 to 0.58 ± 0.14 . Similar patterns were observed for GPT-4.1 and Gemma-4-12B-it.

Table 2
Automated Fidelity Metrics for LLM-Generated Spanish Translations Using Clinician-Framed and Literal Prompting Strategies

Metric	GPT-5 (basic)	GPT-5 (minimal)	GPT-4.1 (basic)	GPT-4.1 (minimal)	Gemma (basic)	Gemma (minimal)	Wilcoxon p-value
BLEU	0.27 ± 0.11	0.38 ± 0.13	0.19 ± 0.09	0.42 ± 0.15	0.23 ± 0.11	0.40 ± 0.15	< 0.001
WER	0.73 ± 0.14	0.58 ± 0.14	0.90 ± 0.18	0.71 ± 0.28	0.76 ± 0.17	0.55 ± 0.31	< 0.001
NIST	4.01 ± 1.04	4.90 ± 1.13	3.48 ± 0.94	5.00 ± 1.33	3.91 ± 1.15	5.20 ± 1.31	< 0.001
ROUGE-L	0.61 ± 0.09	0.70 ± 0.09	0.53 ± 0.09	0.68 ± 0.10	0.57 ± 0.10	0.68 ± 0.11	< 0.001
CUI F1	0.77 ± 0.07	0.82 ± 0.06	0.76 ± 0.06	0.83 ± 0.06	0.82 ± 0.06	0.87 ± 0.05	< 0.001
<p>Values are presented as mean $\pm$ standard deviation across 200 discharge instructions. Automated metrics were computed by comparing English back-translations of LLM-generated Spanish translations with the original English discharge instructions. The basic prompt refers to a clinician-framed translation prompt, whereas the minimal prompt refers to a minimal literal translation prompt. Wilcoxon signed-rank tests compared paired metric values between prompting strategies across all models.</p> <p>Abbreviations: GPT, Generative Pre-trained Transformer; BLEU, Bilingual Evaluation Understudy; WER, Word Error Rate; NIST, National Institute of Standards and Technology metric; ROUGE-L, Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence; CUI F1, Concept Unique Identifier F1 score.</p>							

MQM-Based LLM-as-a-Judge Evaluation

MQM-based evaluation demonstrated variation in domain scores across models and prompting conditions (Fig. 2). Accuracy and terminology were the lowest-scoring domains across all six model–prompt combinations. Accuracy scores ranged from 1.59 ± 1.22 (Gemma-4-12B-it, basic prompt) to 2.23 ± 1.57 (GPT-5, minimal prompt), whereas terminology scores ranged from 3.10 ± 1.53 (GPT-5, minimal prompt) to 3.79 ± 1.46 (GPT-5, basic prompt). In contrast, local convention, audience appropriateness, and design & markup consistently received the highest scores across all models and prompting

strategies, with mean values exceeding 4.18. Fluency and style varied between prompting strategies across all three models. For GPT-5, fluency scores were 4.20 (good) $\pm$ 1.21 for the basic prompt and 3.27 $\pm$ 1.33 for the minimal prompt, while style scores were 4.40 $\pm$ 0.89 and 3.83 $\pm$ 1.21, respectively. Similar differences were observed for GPT-4.1 (fluency: 4.38 vs. 3.73; style: 4.52 vs. 3.94) and Gemma-4-12B-it (fluency: 4.30 vs. 3.27; style: 4.37 vs. 3.79).

Agreement Between LLM-Based and Clinician MQM Evaluation

Agreement between MQM-based LLM-as-a-judge scores and clinician assessments was low across all MQM domains (Table 3). Cohen’s kappa values ranged from -0.18 to 0.20 across both clinician reviewers. The highest agreement between LLM-based evaluation and Clinician Reviewer 1 was observed for audience appropriateness ($\kappa = 0.20$), whereas agreement with Clinician Reviewer 2 was highest for audience appropriateness ($\kappa = 0.13$). Linguistic convention, style, and local convention demonstrated near-zero or negative kappa values for one or both reviewers. Agreement between the two clinician reviewers ranged from $\kappa = -0.11$ to $\kappa = 0.49$ across MQM domains. The highest inter-rater agreement was observed for design & markup ($\kappa = 0.49$), followed by terminology ($\kappa = 0.30$). Accuracy demonstrated the lowest agreement between clinician reviewers ($\kappa = -0.11$).

Table 3
Agreement Between MQM-Based LLM-as-a-Judge Evaluation and Bilingual Clinician Assessment Across MQM Domains

Domain	LLM vs. Clinician 1	LLM vs. Clinician 2	Clinician 1 vs. Clinician 2
Accuracy	0.14	-0.06	-0.11
Terminology	0.09	-0.03	0.30
Linguistic Convention	-0.12	0.01	0.21
Style	0.06	-0.18	0.02
Local Convention	-0.06	-0.03	0.02
Audience Appropriateness	0.20	0.13	0.20
Design and Markup	0.13	-0.11	0.49
<i>Cohen’s kappa coefficients quantify agreement beyond chance between MQM-based LLM evaluation and independent bilingual clinician reviewers for each MQM domain. Values range from -1 to 1, where higher values indicate stronger agreement, 0 indicates agreement equivalent to chance, and negative values indicate less agreement than expected by chance. Analyses were performed on a randomly selected subset of 20 translated discharge instructions.</i>			
<i>Abbreviations: LLM, Large Language Model; MQM, Multidimensional Quality Metric.</i>			

DISCUSSION

In this methodological evaluation of LLM-generated Spanish discharge instruction translations, we found that automated fidelity metrics, MQM-based LLM evaluation, and clinician review measured complementary rather than interchangeable dimensions of translation quality. Although automated metrics favored literal prompting and MQM-based evaluation identified differences across structured quality domains, agreement with clinician assessment remained limited in several domains, particularly those requiring interpretation of clinical meaning.

Automated fidelity metrics consistently favored the literal prompting strategy across all three evaluated models, whereas MQM-based evaluation and clinician review emphasized different aspects of translation quality. Automated metrics such as BLEU, ROUGE-L, and NIST primarily assess lexical overlap and structural similarity between translated and reference text, making them well suited for scalable benchmarking and model comparison. Prior work in machine translation has shown that such metrics do not consistently align with human judgment, especially for tasks that require preserving meaning or contextual interpretation.^{7,8} In clinical settings, this limitation becomes more consequential, as even small shifts in meaning may affect patient understanding and downstream decision-making.²⁰

Agreement between MQM-based LLM evaluation and clinician assessment varied across MQM domains, suggesting that agreement depended on the type of judgment being made. Agreement was highest for local convention, a domain governed largely by standardized linguistic rules, but was lower for accuracy, terminology, and audience appropriateness, which require greater contextual interpretation. This pattern is consistent with prior studies showing that LLMs demonstrate higher agreement with human evaluators on more objective assessment tasks than on evaluations requiring contextual reasoning or subjective interpretation.^{21,22} MQM was originally developed as a framework for structured human evaluation, and prior work has shown that agreement across MQM domains varies according to the type of error being assessed and the experience of the evaluator.²³ Our findings suggest that MQM remains a useful framework for organizing translation quality, but its application through an LLM-as-a-judge may not fully reflect how clinicians evaluate patient-facing communication.

Although MQM-based evaluation produced consistently high scores across models and domains, agreement with clinician assessment remained limited in several MQM domains. This suggests that strong performance within a structured evaluation framework does not necessarily translate to agreement with clinician judgment. Similar observations have been reported in prior studies of machine translation in clinical settings, where outputs are generally usable yet still contain errors that may be clinically relevant.^{5,6,10} For patient-facing discharge instructions, these differences are important because inaccuracies in wording or terminology may affect patient understanding even when overall translation quality appears high. Prior studies among patients with LEP have shown that language discordance is associated with lower comprehension, reduced adherence, and a higher risk of adverse outcomes.^{2,3,24}

From a practical standpoint, our findings suggest that no single evaluation framework is sufficient to assess patient-facing clinical translations. Automated fidelity metrics provide an efficient approach for

benchmarking models and identifying broad differences in translation performance, whereas MQM offers a structured framework for evaluating specific domains of translation quality. However, neither approach consistently reflected clinician judgment in domains requiring contextual interpretation. Human review therefore remains an important component of translation evaluation, particularly for assessing accuracy, terminology, and other domains where clinically meaningful differences are most likely to occur. Rather than serving as competing approaches, automated fidelity metrics, MQM-based evaluation, and clinician review each provide complementary information that can inform a more comprehensive assessment of translation quality. This approach aligns with broader work suggesting that human evaluation remains central in high-stakes LLM-assisted tasks.^{12,25,26}

Limitations

This study has several limitations. Professional human translations were not available as an external translation benchmark. Clinician validation was performed on a subset of translations and relied on two bilingual reviewers with different experience utilizing MQM. Consequently, agreement estimates should be interpreted as preliminary and hypothesis-generating. Future studies should include multiple independent reviewers and adjudication procedures to establish the reproducibility of clinician-based MQM assessment. We also did not directly assess patient comprehension or usability, both of which are important endpoints for discharge communication. In addition, this analysis focused on a single language pair and clinical context, which may limit generalizability to other settings or types of clinical text. Future work should incorporate patient-centered outcomes, including comprehension and usability, alongside clinician and automated evaluation. There may also be value in refining LLM-as-judge approaches to better align with human interpretation, particularly in domains where context and meaning play a larger role

CONCLUSION

Evaluation of LLM-generated clinical translations requires multiple complementary approaches because different evaluation frameworks capture different dimensions of translation quality. Automated fidelity metrics, MQM-based evaluation, and clinician review each contribute unique information that cannot be obtained from a single method alone. A layered evaluation strategy may provide a more comprehensive framework for evaluating patient-facing clinical translations before clinical implementation.

DECLARATIONS

COMPETING INTEREST

AL is co-founder of Atlas Care AI, Inc and holds equity in Glass Health.

RESPONSIBLE USE OF AI DISCLOSURE

The authors used generative AI tools, including ChatGPT (OpenAI), and writing assistance software (Grammarly) to support manuscript editing, structural organization, and language clarity. These tools were not used to generate original data, analyses, or primary scientific content. All scientific content, interpretation, and conclusions are the original work of the authors, who take full responsibility for the manuscript.

Author Contribution

W.M. conceptualized the study, drafted the manuscript, and led the overall project. S.Y. and Y.G. conducted data analysis and contributed to interpreting the results. E.G. contributed to study design and analytic methods and provided technical and methodological guidance. Y.G. provided senior oversight, contributed to study design and interpretation, and supervised the project. All authors reviewed, revised, and approved the final manuscript.

Acknowledgement

The authors thank Maya Kruse, MS, and Morgan Pena, MPH, for their contributions to the early stages of the project. The authors thank Carolina Gomez Grimaldi, MD for her contribution to bilingual clinician MQM-based scoring. This work was partially supported by the National Institutes of Health / National Library of Medicine (NIH/NLM) under award number R00LM14308 and NIH/NCATS Colorado CTSA Grant Number UM1 TR004399. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

DATA AVAILABILITY

The data used in this study were derived from the Medical Information Mart for Intensive Care (MIMIC) database, which is publicly available to qualified researchers upon completion of required training and data use agreements (<https://physionet.org/>). Derived data supporting the findings of this study, including translated outputs and evaluation results, are available from the corresponding author upon reasonable request, subject to institutional and data use restrictions. The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

REFERENCES

1. Austad, K. *et al.* Evaluating the quality and equity of patient hospital discharge instructions. *BMC Health Serv. Res.* 25, 291 (2025).

2. Anaya, M., Law, W., Montoya, H. L. & Moreira, C. M. Discharge Interventions for Limited English Proficiency Patients: A Scoping Review. *J. Racial Ethn. Health Disparities* <https://doi.org/10.1007/s40615-024-02003-2> (2024) doi:10.1007/s40615-024-02003-2.
3. Rawal, S. et al. Association Between Limited English Proficiency and Revisits and Readmissions After Hospitalization for Patients With Acute and Chronic Conditions in Toronto, Ontario, Canada. *JAMA* 322, 1605 (2019).
4. Davis, S. H. et al. Translating Discharge Instructions for Limited English-Proficient Families: Strategies and Barriers. *Hosp. Pediatr.* 9, 779–787 (2019).
5. Kong, M. et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. *BMJ Qual. Saf.* bmjqs-2024-018384 (2025) doi:10.1136/bmjqs-2024-018384.
6. Brewster, R. C. L. et al. Performance of ChatGPT and Google Translate for Pediatric Discharge Instruction Translation. *Pediatrics* 154, e2023065573 (2024).
7. Rei, R., Stewart, C., Farinha, A. C. & Lavie, A. COMET: A Neural Framework for MT Evaluation. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2685–2702 (Association for Computational Linguistics, Online, 2020). doi:10.18653/v1/2020.emnlp-main.213.
8. Glushkova, T., Zerva, C. & Martins, A. F. T. BLEU Meets COMET: Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation. Preprint at <https://doi.org/10.48550/ARXIV.2305.19144> (2023).
9. Lommel, A., Uszkoreit, H. & Burchardt, A. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica* 0455–0463 (2014).
10. Ray, M. et al. Evaluating a Large Language Model in Translating Patient Instructions to Spanish Using a Standardized Framework. *JAMA Pediatr.* <https://doi.org/10.1001/jamapediatrics.2025.1729> (2025) doi:10.1001/jamapediatrics.2025.1729.
11. Muhanna, A. A. I. Assessing machine translation quality on uterine cancer content: MQM-based comparative study. *CDELTA Occas. Pap. Dev. Engl. Educ.* 89, 421–458 (2025).
12. Croxford, E. et al. Current and future state of evaluation of large language models for medical summarization tasks. *Npj Health Syst.* 2, 6 (2025).
13. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. in 311–318 (2002).
14. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. in 74–81 (2004).
15. Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. in 138–145 (2002).
16. Klakow, D. & Peters, J. Testing the correlation of word error rate and perplexity. *Speech Commun.* 38, 19–28 (2002).

17. Gao, Y. *et al.* Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. in 2979–2991 (2022).
18. Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, D267–D270 (2004).
19. Soldaini, L. & Goharian, N. Quickumls: a fast, unsupervised approach for medical concept extraction. in 1–4 (2016).
20. Han, L. *et al.* Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. *Front. Digit. Health* 6, 1211564 (2024).
21. Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* 36, 46595–46623 (2023).
22. Huang, H. *et al.* An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. in 5880–5895 (2025).
23. Snow, T. A. *Establishing the Viability of the Multidimensional Quality Metrics Framework.* (Brigham Young University, 2015).
24. Khoong, E. C., Steinbrook, E., Brown, C. & Fernandez, A. Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. *JAMA Intern. Med.* 179, 580–582 (2019).
25. Zhao, X. *et al.* Rx-LLM: a benchmarking suite to evaluate safe large language model performance for medication-related tasks. *MedRxiv Prepr. Serv. Health Sci.* 2025.12.01.25341004 (2025) doi:10.64898/2025.12.01.25341004.
26. Gao, Y. *et al.* A scoping review of publicly available language tasks in clinical natural language processing. *J. Am. Med. Inform. Assoc.* 29, 1797–1806 (2022).

Figures

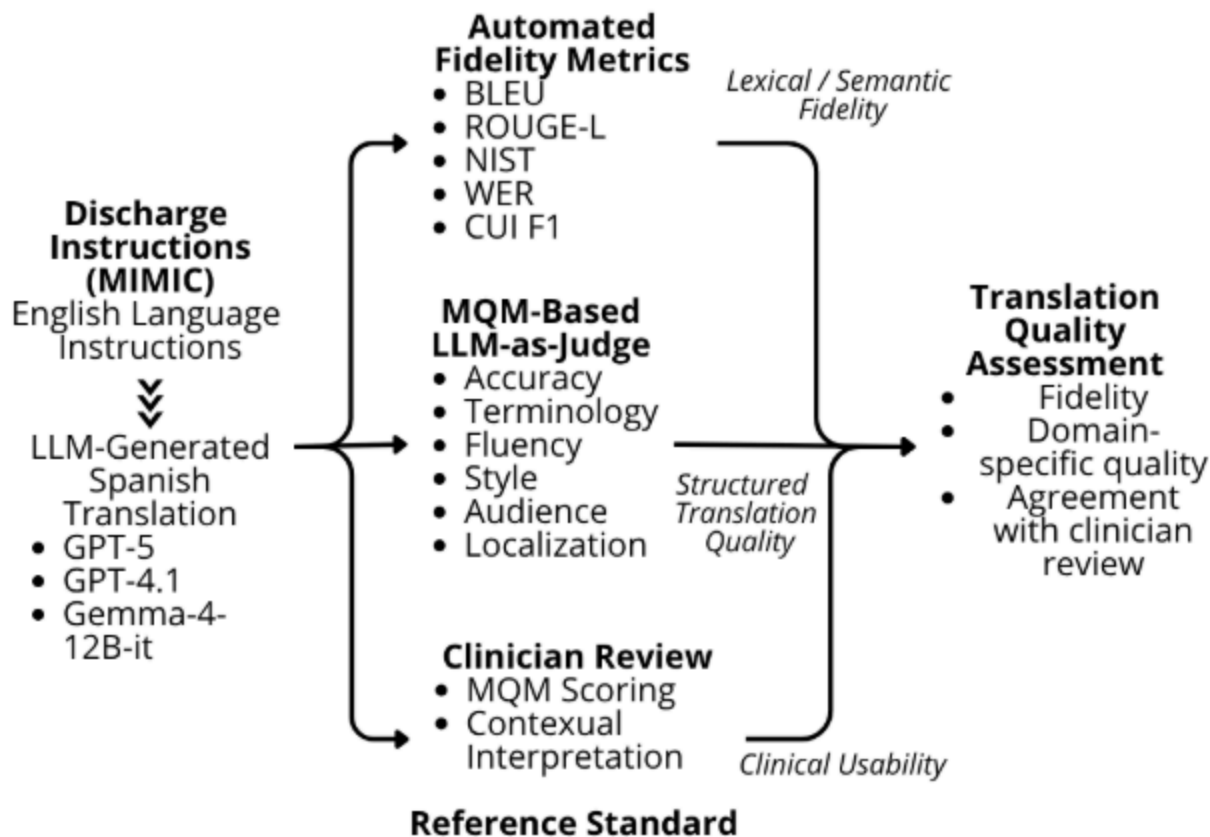


Figure 1

Multi-method framework for evaluating LLM-generated Spanish discharge instruction translations. English-language discharge instructions were translated into Spanish using multiple LLMs and evaluated using automated fidelity metrics, MQM-based LLM-as-a-judge scoring, and bilingual clinician review. Automated metrics quantify lexical and semantic fidelity, MQM-based evaluation provides structured domain-level assessment, and clinician review serves as the reference standard for evaluating contextual meaning and clinical usability. Agreement across these complementary evaluation approaches provides a comprehensive assessment of translation quality. Figure created with Canva: <https://canva.link/oin8y352986rnum>

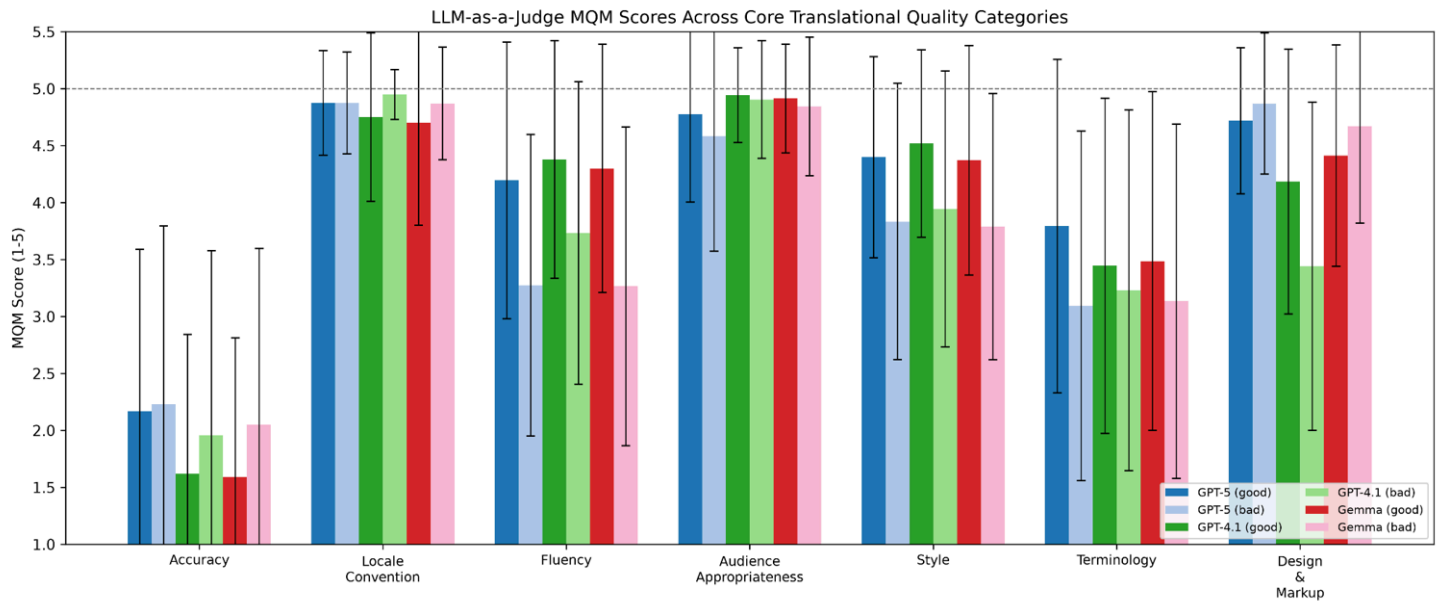


Figure 2

MQM-based LLM-as-a-judge scores across translation quality domains by model and prompting strategy. Mean MQM scores ($\pm$ standard deviation) are shown for seven translation quality domains across GPT-5, GPT-4.1, and Gemma-4-12B-it using clinician-framed and literal prompting strategies. Higher scores indicate better translation quality (5 = no error; 1 = critical error). MQM evaluation demonstrated substantial variation across quality domains, with consistently high scores for local convention, audience appropriateness, and design & markup, and lower scores for accuracy and terminology. Prompting strategy primarily influenced fluency, style, and audience appropriateness, whereas accuracy and terminology remained comparatively challenging across all models.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Appendix.docx](#)