Research Letter

# Evaluating Spanish Translations of Emergency Department Discharge Instructions by a Large Language Model: Tool Validation and Reliability Study

Jossie A Carreras Tartak[1], MD, MBA; Ryan CL Brewster[2], MD; Daniela Arango Isaza[3], MD; Antonio Berumen Martinez[3], MD; Ana Grafals[1], BS; Phanidhar Adusumilli[4], BE; Ted Fitzgerald[4], BS; Roger Orcutt[1], MBA; Larry A Nathanson[1], MD; Adrian D Haimovich[1], MD, PhD

[1]Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States

[2]Department of Pediatrics, Beth Israel Deaconess Medical Center, Boston, MA, United States

[3]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States

[4]Department of Technology and Innovation, Beth Israel Lahey Health, Boston, MA, United States

**Corresponding Author:**
Jossie A Carreras Tartak, MD, MBA
Department of Emergency Medicine
Beth Israel Deaconess Medical Center
1 Deaconess Rd
Boston, MA, 02215
United States
Phone: 1 617 754 2339
Email: jcarrera@bidmc.harvard.edu

## Abstract

When given a sample of 100 emergency department discharge instructions, Claude Sonnet, a large language model, produced accurate Spanish translations as evaluated by Spanish-speaking physicians and medical interpreters.

***(JMIR Form Res 2026;10:e79676)*** doi: 10.2196/79676

**KEYWORDS**

artificial intelligence; machine learning; machine translation; language; disparities

## Introduction

Language-concordant emergency department (ED) discharge instructions are an essential component of equitable care for patients who prefer a language other than English [1-4]. ED discharge instructions are often complex, combining standardized templates with personalized clinician-written text. In most cases, patients who prefer a language other than English still receive instructions in English [5]. When translation is attempted, clinicians will often informally rely on tools such as Google Translate that are not auditable, are generally not institutionally approved for clinical use, and have known performance limitations for long or technically detailed documents [6-8].

Large language models (LLMs) offer a promising auditable and institutionally governable approach to addressing this equity gap [6,9]. Because reproducibility in patient-care processes requires controlled models rather than open chat interfaces such as ChatGPT (OpenAI), institutions will need to deploy translation LLMs directly [6]. To our knowledge, no formative pilot studies have evaluated LLMs for translating ED discharge instructions. We therefore conducted a feasibility study using real ED discharge instructions to assess LLM translation performance in preparation for clinical implementation.

## Methods

### Study Design

This was a single-center feasibility study at an urban academic medical center. We iteratively developed a translation prompt using Claude Sonnet (version 3.5; Anthropic) accessed via protected health information (PHI)–compliant Amazon Web Services (Amazon), testing each version on batches of 10 to 20 randomly sampled free-text discharge instructions provided to ED patients between July 1 and December 31, 2024. Claude Sonnet 3.5 was selected as it balances cost, performance, and speed and was available in our PHI-compliant environment.

Following prompt development, a team of independent evaluators (2 native Spanish-speaking physicians and 2 certified medical interpreters) reviewed a translated set of 100 randomly sampled free-text discharge instructions. We used a rubric adapted from prior studies consisting of a 5-point Likert scale across 5 domains (Multimedia Appendix 1) designed so that items rated a 3 or lower would be deemed substandard or unacceptable for use in a clinical setting [9]. Scores of 3 or lower in any domain required written explanation and were escalated for further review.

Our primary outcome was the proportion of discharge instructions that scored a 3 or lower in any one domain. The secondary outcome was the mean Likert score for each of the 5 domains stratified by reviewer type (interpreter vs physician). Descriptive analyses were performed in Python (version 3.11).

### Ethical Considerations

This study was approved by the Beth Israel Deaconess Institutional Review Board (2024P000315). All abstracted data were deidentified.

## Results

Of the 100 samples translated using the designed prompt (Multimedia Appendix 2), the mean Likert score ratings across samples by domain were as follows: 5.0 (95% CI 5.0-5.0) for completeness, 4.8 (95% CI 4.8-4.8) for fluency, 4.9 (95% CI 4.9-4.9) for meaning, 5.0 (95% CI 5.0-5.0) for severity, and 4.9 (95% CI 4.9-4.9) overall (Table 1; example translations are provided in Multimedia Appendices 3 and 4). One sample was given a score of 3 by a single reviewer in the domains of meaning and overall quality because the term “concussion” was translated as *conmoción cerebral* (full redacted translation in Multimedia Appendix 3). On adjudication, the translation was deemed clinically acceptable because the term *conmoción cerebral* is one of several translations of the term “concussion,” along with *concusión*.

**Table 1.** Interpreter and physician evaluator scores for Spanish translations (N=100).

| Domain | Mean interpreter scores (95% CI) | Mean physician scores (95% CI) |
|---|---|---|
| Completeness | 5.0 (5.0-5.0) | 5.0 (5.0-5.0) |
| Fluency | 4.8 (4.7-4.9) | 4.8 (4.7-4.9) |
| Meaning | 5.0 (5.0-5.0) | 4.9 (4.9-4.9) |
| Severity | 5.0 (5.0-5.0) | 5.0 (5.0-5.0) |
| Overall | 5.0 (5.0-5.0) | 4.9 (4.9-4.9) |

## Discussion

In this feasibility pilot study, we found that Claude Sonnet produced clinically acceptable Spanish translations of ED discharge instructions. The one case flagged for further review reflected regional differences in Spanish vocabulary, an observation suggesting that future LLM prompts may incorporate patient nationality or dialects to improve comprehensibility.

Our results are in alignment with prior work on standardized discharge instructions as well as free-text instructions from pediatric settings [9,10]. Free-text instructions have the potential for grammatical errors, dictation and typographical errors, missing information, formatting issues, and use of overly complicated medical terminology that might compromise translation quality. A recent study (n=20) in the pediatric setting showed comparable quality between interpreter translation and the GPT-4o model from OpenAI [10]. Our study did not directly compare the LLM outputs to interpreter outputs, but instead included interpreters as reviewers.

Our single-center results may not apply to institutions that have different discharge instruction processes or lack access to PHI-compliant LLMs. Moreover, our study was limited to Spanish. Further testing will be needed to establish the safety of LLM translation before live implementation.

### Acknowledgments

We would like to thank Shari Gold-Gomez, Ana Torres, Natalia Chilcote, and Marie Rodriguez for their contributions to this study. RCLB was affiliated with the Department of Pediatrics at Beth Israel Deaconess Medical Center at the time of the study and is currently affiliated with the Department of Pediatrics at Stanford University School of Medicine. DAI and ABM were affiliated with the Department of Medicine at Beth Israel Deaconess Medical Center at the time of the study. DAI is currently affiliated with the Department of Medicine at the University of California, San Francisco. ABM is currently affiliated with the Department of Medicine at Boston Medical Center.

### Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

## Authors' Contributions

Conceptualization: ADH (lead), JACT (equal), RCLB (supporting)
Data curation: ADH
Formal analysis: ADH
Funding acquisition: ADH (lead), JACT (equal)
Investigation: DAI (lead), ABM (equal)
Methodology: ADH (lead), JACT (equal), AG (supporting), RCLB (supporting)
Project administration: ADH (lead), JACT (equal)
Resources: LAN (lead), RO (supporting)
Software: TF (lead), PA (supporting)
Supervision: ADH (lead), JACT (equal)
Writing—original draft: JACT
Writing—review and editing: JACT (lead), ADH (supporting)

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Interpretation rating guide.
[DOCX File , 17 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

Large language model prompt.
[DOCX File , 15 KB-Multimedia Appendix 2]

## Multimedia Appendix 3

Translation with low score.
[DOCX File , 18 KB-Multimedia Appendix 3]

## Multimedia Appendix 4

Sample translation.
[DOCX File , 16 KB-Multimedia Appendix 4]

## References

1. Khoong EC, Sherwin EB, Harrison JD, Wheeler M, Shah SJ, Mourad M, et al. Impact of standardized, language-concordant hospital discharge instructions on postdischarge medication questions. J Hosp Med. Sep 2023;18(9):822-828. [FREE Full text] [doi: 10.1002/jhm.13172] [Medline: 37490045]
2. Samuels-Kalow ME, Stack AM, Porter SC. Effective discharge communication in the emergency department. Ann Emerg Med. Aug 2012;60(2):152-159. [doi: 10.1016/j.annemergmed.2011.10.023] [Medline: 22221840]
3. Gutman CK, Lion KC, Fisher CL, Aronson PL, Patterson M, Fernandez R. Breaking through barriers: the need for effective research to promote language-concordant communication as a facilitator of equitable emergency care. J Am Coll Emerg Physicians Open. Feb 2022;3(1):e12639. [FREE Full text] [doi: 10.1002/emp2.12639] [Medline: 35072163]
4. Lion KC, Lin Y, Kim T. Artificial intelligence for language translation: the equity is in the details. JAMA. Nov 05, 2024;332(17):1427-1428. [doi: 10.1001/jama.2024.15296] [Medline: 39264601]
5. Isbey S, Badolato G, Kline J. Pediatric emergency department discharge instructions for Spanish-speaking families: are we getting it right? Pediatr Emerg Care. Feb 01, 2022;38(2):e867-e870. [doi: 10.1097/PEC.0000000000002470] [Medline: 34140448]
6. Lopez I, Velasquez DE, Chen JH, Rodriguez JA. Operationalizing machine-assisted translation in healthcare. NPJ Digit Med. Sep 30, 2025;8(1):584. [FREE Full text] [doi: 10.1038/s41746-025-01944-0] [Medline: 41028827]
7. Khoong EC, 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. Apr 01, 2019;179(4):580-582. [FREE Full text] [doi: 10.1001/jamainternmed.2018.7653] [Medline: 30801626]
8. Taira BR, Kreger V, Orue A, Diamond LC. A pragmatic assessment of Google Translate for emergency department instructions. J Gen Intern Med. Nov 2021;36(11):3361-3365. [FREE Full text] [doi: 10.1007/s11606-021-06666-z] [Medline: 33674922]

9. Ray M, Kats DJ, Moorkens J, Rai D, Shaar N, Quinones D, et al. Evaluating a large language model in translating patient instructions to Spanish using a standardized framework. JAMA Pediatr. Sep 01, 2025;179(9):1026-1033. [doi: 10.1001/jamapediatrics.2025.1729] [Medline: 40622720]
10. Brewster RCL, Gonzalez P, Khazanchi R, Butler A, Selcer R, Chu D, et al. Performance of ChatGPT and Google Translate for pediatric discharge instruction translation. Pediatrics. Jul 01, 2024;154(1):e2023065573. [doi: 10.1542/peds.2023-065573] [Medline: 38860299]

## Abbreviations

**ED:** emergency department
**LLM:** large language model
**PHI:** protected health information

*Edited by A Stone; submitted 26.Jun.2025; peer-reviewed by PB Chandrashekar, R Frederking, UK Chilakalapalli; comments to author 01.Oct.2025; revised version received 13.Nov.2025; accepted 13.Nov.2025; published 12.Jan.2026*

*Please cite as:*
*Carreras Tartak JA, Brewster RCL, Arango Isaza D, Berumen Martinez A, Grafals A, Adusumilli P, Fitzgerald T, Orcutt R, Nathanson LA, Haimovich AD*
*Evaluating Spanish Translations of Emergency Department Discharge Instructions by a Large Language Model: Tool Validation and Reliability Study*
*JMIR Form Res 2026;10:e79676*
*URL: https://formative.jmir.org/2026/1/e79676*
*doi: 10.2196/79676*
*PMID:*