# Transcending Language Barriers: Evaluation of GPT-4's Translation Accuracy in Dermatological Patient Education Materials

Elyse Mackenzie, BS[1], Sydney Wolfe, BA[1], Bianca Sanabria, MA[1.2]. Contributing author: Kimberly E Siehl, MSW, LSW[3]

[1]Rutgers Robert Wood Johnson Medical School, Piscataway, NJ; [2]Center for Dermatology, Rutgers Robert Wood Johnson Medical School, Somerset, NJ; [3]RWJ Barnabas Health- Jersey City Medical Center.

## Abstract

This pilot study **evaluates GPT-4's capability to translate English Patient Education Materials (PEMs) into Spanish**, with aims **to reduce language barriers in healthcare**. Focusing on dermatitis, PEMs were analyzed for readability and accuracy by a bilingual rater and traditional readability scales.

GPT-4 translations scored well in conveying accurate information (Adequacy, Meaning, Severity) with slightly lower Fluency as compared to the original translation. Despite limitations like a small sample size and single rater analysis, the results indicate that **AI could significantly aid in creating accessible health materials for non-English speaking populations,** potentially impacting global health education. Future research will broaden evaluation parameters and test with end-users to validate these findings.

## Background

- Culturally and literacy-appropriate **Patient Education Materials (PEMs)** are essential to address health disparities, especially where follow-up care is limited.[1]
- Most U.S. healthcare PEMs are in English, **disadvantaging non-English speakers** and potentially leading to worse health outcomes.[2]
- There is **a lack of research on the readability of non-English PEMs,** raising concerns for the large Spanish-speaking population in the U.S.[3]
- **Skin diseases have a significant global health impact,** highlighting the need for adequate multilingual dermatological PEMs.[4]
- Artificial intelligence tools like **ChatGPT may offer an accessible, cost-effective, and efficient alternative** to human translators, but its translation capacity has **not yet been** tested in dermatology.[5]

## References

1. Wittink H, Oosterhaven J. Patient education and health literacy. Musculoskelet Sci Pract. 2018;38:120-127. doi:10.1016/j.msksp.2018.06.004
2. Harvey I, O'Brien M. Addressing health disparities through patient education: the development of culturally-tailored health education materials at Puentes de Salud. J Community Health Nurs. 2011;28(4):181-189. doi:10.1080/07370016.2011.614827
3. 4. Patetta MJ, Pond KM, Tennant EM, Sood A, Gonzalez MH. Readability Level of English and Spanish Orthopaedic Patient Education Materials English and Spanish Patient Education. J Surg Orthop Adv. 2021;30(2):96-100.
4. Yakupu A, Aimaier R, Yuan B, et al. The burden of skin and subcutaneous diseases: findings from the global burden of disease study 2019. Frontiers in Public Health. 2023;11. Accessed November 19, 2023. https://www.frontiersin.org/articles/10.3389/fpubh.2023.1145513
5. Introducing ChatGPT. Accessed July 28, 2023. https://openai.com/blog/chatgpt
6. Yakupu A, Aimaier R, Yuan B, et al. The burden of skin and subcutaneous diseases: findings from the global burden of disease study 2019. Frontiers in Public Health. 2023;11. Accessed November 19, 2023. https://www.frontiersin.org/articles/10.3389/fpubh.2023.1145513
7. Khanna RR, Karliner LS, Eck M, Vittinghoff E, Koenig CJ, Fang MC. Performance of an online translation tool when applied to patient educational material. J Hosp Med. 2011;6(9):519-525. doi:10.1002/jhm.898
8. Fernández AM. Analizador de legibilidad de texto. Legible. Published November 27, 2016. Accessed November 20, 2023. https://legible.es/

## Methods and Materials

- Data on skin diseases was sourced from the **2019 Global Burden of Disease Study,** focusing on dermatitis as the most burdensome condition.[6]
- Three **PEMs about atopic dermatitis from JAMA Dermatology** were analyzed in both English and Spanish.
- GPT-4 translated the English PEMs into Spanish.
- A **bilingual social worker** evaluated GPT and original translations on a Likert scale for fluency, adequacy, meaning, severity, and preference.
- Readability was assessed using **Flesch-Kincaid Reading Ease for English and Fernández-Huerta scores** for Spanish translations.[7,8] (Figure 2)

**Figure 1: Sample GPT Prompt and Output**

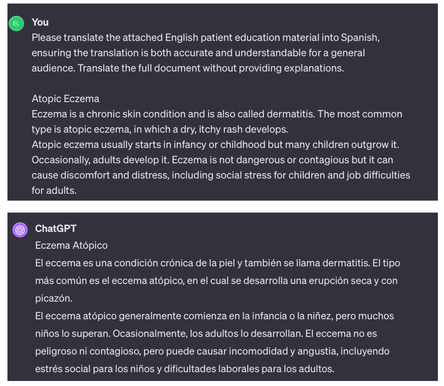
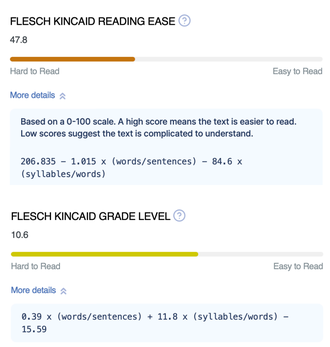



**Figure 2: Sample Flesch-Kincaid and Fernández Huerta Outputs[7,8]**




Legibilidad del texto

| índice | valor | dificultad |
|---|---|---|
| Fernández Huerta | 50.94 | algo difícil |
| Gutiérrez | 37.66 | normal |
| Szigriszt-Pazos | 45.83 | bastante difícil |
| INFLESZ | 45.83 | algo difícil |
| legibilidad $\mu$ | 54.21 | un poco difícil |

## Results

**Table I:** Dermatological Burden of Disease in Disability Adjusted Life Years (DALYs), 2019: Top 3 by Region[a]

| Global | United States | Southern Latin America |
|---|---|---|
| 1. Dermatitis; 10,063,791.45 [5,956,892.77, 15,750,929.61] | 1. Dermatitis; 555,327.74 [333,786.17, 857,463.56] | 1. Dermatitis; 88,915.27 [49,844.04, 145,593.82] |
| 2. Acne vulgaris; 4,955,197.67 [2,978,062.72, 7,853,723.69] | 2. Viral skin diseases; 391,770.99 [255,564.88, 578,002.78] | 2. Viral skin diseases; 73,536.66 [47,606.86, 109,430.68] |
| 3. Scabies; 4,837,980.63 [2,682,199.12, 7,727,672.44] | 3. Psoriasis; 353,064.98 [255,001.87, 464,079.87] | 3. Psoriasis; 56,128.17 [39,341.30, 74,462.40] |
| 4. Viral skin diseases; 4,728,083.85 [3,026,072.11, 7,062,875.76] | 4. Malignant skin melanoma; 308,801.00 [240,818.39, 419,196.24] | 4. Acne vulgaris; 47,835.02 [28,635.91, 77,062.79] |
| 5. Urticaria; 3,898,838.56 [2,554,225.43, 5,584,365.55] | 5. Other skin and subcutaneous diseases; 252,247.96 [125,263.60, 458,527.07] | 5. Other skin and subcutaneous diseases; 42,514.23 [21,392.86, 76,369.88] |

[a]Disability adjusted life years (DALYs) are reported in numbers with their 95% confidence interval, where one DALY represents the loss of one year of full health. DALYs are the sum of years of life lost due to premature mortality (YLLs) and years lived with a disability (YLDs) due to the disease.[7]

**Table II.** Score Comparison by Translation Method

| | Article A | | Article B | | Article C | | Average Scores |
|---|---|---|---|---|---|---|---|
| | GPT-4 | Original | GPT-4 | Original | GPT-4 | Original | GPT / Original |
| Fluency | 4 | 5 | 4 | 5 | 4 | 4 | 4 / 4.67 |
| Adequacy | 5 | 5 | 5 | 5 | 5 | 5 | 5 / 5 |
| Meaning | 5 | 5 | 5 | 5 | 5 | 5 | 5 / 5 |
| Severity | 5 | 5 | 5 | 5 | 5 | 5 | 5 / 5 |
| Preference | 0 | 5 | 0 | 5 | 5 | 0 | 5 / 10 |

**Table III.** Readability Scores

| **Readability: JAMA Original PEMS (English)** | | | | |
|---|---|---|---|---|
| | Article A | Article B | Article C | Average |
| Flesch Kincaid Reading Ease | 10.3 | 58.3 | 31 | 33.2 |
| Flesch Kincaid Grade Level | 18 | 8.5 | 14.5 | 13.67 |
| **Readability: GPT-4 Spanish Translation** | | | | |
| | Article A | Article B | Article C | Average |
| Fernandez Huerta Score | 48.74 | 61.43 | 41.79 | 50.65 |
| **Readability: JAMA Original Spanish Translation** | | | | |
| Fernandez Huerta Score | 49.9 | 62.58 | 45.81 | 52.76 |

- Original translations outperformed GPT-4 in Fluency scores, with originals scoring an average of 4.67 compared to GPT-4's 4, and raters showing a preference for originals in most cases.
- The original English articles were found to be quite difficult to read, with a Flesch-Kincaid Reading Ease score averaging 33.2 and a corresponding grade level of 13.67.
- Original Spanish translations were slightly more readable than GPT-4 translations, scoring an average of 52.76 on the Fernández-Huerta scale compared to GPT-4's 50.65, indicating that both are "fairly difficult to read" but the original is marginally easier.

## Conclusions & Future Directions

The study suggests **AI translation tools could revolutionize patient education globally**, despite slight readability challenges.

GPT-4 and original Spanish translations both received the top scores for Adequacy, Meaning, and Severity, indicating **high-quality translation that does not adversely affect patient health outcomes**. GPT-4's translations were slightly less fluent than professional ones, yet one was preferred over the original in terms of usability. English PEMs were objectively complex with reading levels much higher than recommended, but Spanish translations maintained similar reading levels, suggesting that translations mimic original reading levels.

Overall, these scores reflect GPT-4's **significant potential in translating PEMs from English to Spanish.**

Future studies may increase the sample size and involve multiple bilingual raters to enhance the power and generalizability of the study's results. Studies might also test the effectiveness of AI-translated PEMs directly with patients to evaluate real-world usability.