Annals of Case Reports
Cohen LE, et al. Ann Case Rep: 10: 102376
www.doi.org/10.29011/2574-7754.102376
www.gavinpublishers.com

OPEN ACCESS

GAVIN PUBLISHERS

Research Article

Open AI Chat GPT as a Patient-Facing Translator: Evaluating Artificial Intelligence-Generated Spanish and English Summaries of Breast Imaging Reports

Lucas E. Cohen[1*], Kiran N. Chang[2]

[1]The Kinkaid School, Houston Texas, USA

[2]Department of Diagnostic and Interventional Imaging, UT Health McGovern Medical School, Houston, Texas, USA

[*]**Corresponding author**: Lucas E. Cohen, The Kinkaid School, Houston Texas, 201 Kinkaid School Drive, Houston, Texas, 77024, USA

Citation: Cohen LE, Chang KN (2025) Open AI Chat GPT as a Patient-Facing Translator: Evaluating Artificial Intelligence-Generated Spanish and English Summaries of Breast Imaging Reports. Ann Case Report. 10: 2376. DOI: 10.29011/2574-7754.102376

Received: 11 August 2025; **Accepted**: 15 August 2025; **Published**: 18 August 2025

Abstract

Objective: Breast imaging reports are typically written for healthcare professionals using technical language that limits patient comprehension. This study evaluated ChatGPT's ability to simplify breast radiology reports into Spanish while maintaining clinical accuracy.

Methods: We collected 50 publicly posted breast imaging reports from Reddit, an online discussion platform, and used GPT-4o (March 2025 version) to generate simplified versions in both English and Spanish. Readability was assessed using Flesch Reading Ease Score (FRES) and Flesch Kincaid Grade Level (FKRL) for English outputs and Fernández-Huerta and Szigriszt-Pazos indices for Spanish outputs. A bilingual, fellowship-trained academic breast radiologist reviewed all outputs for clinical accuracy. Analysis included paired t-tests and Wilcoxon signed-rank tests.

Results: ChatGPT significantly improved English readability, reducing FKRL from 11.6 ± 1.8 to 9.6 ± 1.1 and increasing FRES from 33.9 ± 11.8 to 59.7 ± 5.4 (both $p < .001$). Word count was slightly higher for English outputs (248 ± 80 vs. 240 ± 140 words, $p = .03$), indicating gains were due to language simplification rather than content reduction. Spanish translations achieved Fernández-Huerta and Szigriszt-Pazos scores of 87.0 ± 5.1 and 83.5 ± 5.2, respectively both in the "*easy*" readability range with no outputs exceeding the equivalent of an 8th-grade FKRL score. All English and Spanish outputs (100%) preserved clinical accuracy. Subgroup analysis showed that even the most complex reports, such as those involving biopsy, were effectively simplified without loss of meaning.

Conclusion: ChatGPT enhances the readability of breast imaging reports in English and Spanish while preserving fidelity.

Keywords: Breast Imaging; Radiology Report Simplification; ChatGPT; Artificial Intelligence; Spanish Translation; Health Literacy

Introduction

Radiology reports for mammography and breast ultrasound are typically written for healthcare providers rather than patients, using dense medical terminology, abbreviations, and complex sentence structures that significantly impair lay understanding [1]. This communication barrier is especially concerning in breast imaging, where high patient anxiety often accompanies potential cancer diagnoses [2]. The implementation of the 21st Century Cures Act in the United States has further complicated this dynamic by granting patients immediate access to their radiology results, including breast imaging reports, often before any physician has reviewed the findings with them [3, 4]. Many patients find themselves faced with technically written results, prompting them to search the internet including forums, social media, and patient networks for explanations [5,6]. This phenomenon is visible on platforms like Reddit, an online forum and discussion platform (www.reddit.

com), where users frequently post de-identified breast imaging reports seeking help interpreting their results. These real-world data points reflect authentic patient confusion and information needs [1,7].

Compounding this issue is the language barrier. Spanish is the most spoken non-English language in the United States, with approximately 13% of households speaking Spanish at home representing about 62% of all non-English speakers [8]. Yet, high-quality breast imaging education materials in Spanish remain limited [9]. Furthermore, most English-language breast imaging educational content is written at a 10th to 11th grade reading level well above the 6th grade level recommended by the AMA and NIH for health information materials [10, 11].

Large language models (LLM) like ChatGPT (www.chatGPT.com) offer a promising solution. Studies have demonstrated their capacity to simplify complex clinical text while retaining core meaning, with ChatGPT performing especially well in radiology and breast imaging contexts [12, 13]. Tools like ChatGPT, Gemini, and Microsoft Copilot have also shown the ability to accurately respond to frequently asked breast imaging questions, albeit with variability in readability levels [13]. Prior research has primarily examined readability improvements in English, using curated or synthetic cases, rather than real-world, patient-generated data [14-20]. Critically, no published study to date has evaluated ChatGPT's ability to translate breast imaging reports into Spanish from actual reports posted online by patients—a clear gap given the communication challenges faced by the Spanish-speaking population.

To address this, we evaluated ChatGPT's ability to simplify and translate into Spanish a set of 50 publicly posted breast imaging reports, originally written in English and shared online by patients. The outputs were assessed across three domains: readability, word counts, and clinical accuracy. We hypothesized that ChatGPT would generate Spanish-language outputs that are significantly more readable than the original English reports, while maintaining clinical fidelity—potentially offering a scalable solution to address a widespread communication gap in breast radiology.

Materials and Methods

Study Design and Ethical Approval

This cross-sectional, observational study evaluated the ability of OpenAI's ChatGPT to translate and simplify breast radiology reports, specifically mammography and breast ultrasound, into patient-friendly Spanish and English summaries. The study protocol was deemed exempt by The University of Texas MD Anderson Cancer Center institutional review board, and a waiver of informed consent was granted due to the retrospective design and the use of publicly available, de-identified data.

Data Source and Report Selection

A total of 50 breast imaging reports were collected from Reddit (www.reddit.com), a publicly accessible online platform where patients often post medical content and seek interpretation. Reports were identified by searching for the single keyword "mammogram report." To be included, posts had to contain full-text imaging reports written in English, include a direct request for clarification, and be free of identifiable patient information. Reports that did not meet these criteria were excluded. All reports were anonymized before analysis.

ChatGPT Translation and Simplification

The English text from each publicly posted report was entered into ChatGPT (version GPT-4o, San Francisco, OpenAI) to produce two distinct patient-facing outputs: a simplified English version and a Spanish translation with simplified language. Standardized prompts were used for consistency. For English simplification, the prompt was: "Explain this radiology report to a patient in layman's terms in second person: <Report Text>". For Spanish translation and simplification, the prompt was: "Explica este informe de radiología al paciente en términos simples y en segunda persona: <Report Text>." All ChatGPT outputs were saved for further analysis.

Readability and Word Count Assessment

Word counts and readability metrics were computed for both the original reports and the ChatGPT-generated outputs using Python (www.python.org) with the textstat package (pypi.org/project/textstat). For English texts, the Flesch Reading Ease Score (FRES) and Flesch–Kincaid Grade Level (FKRL) were calculated. For Spanish outputs, the Fernández-Huerta index and INFLESZ (Índice de legibilidad de Szigriszt-Pazos) were applied. All metrics were obtained using standardized textstat functions.

Accuracy and Fidelity Review

Accuracy was defined as the presence or absence of any clinically significant deviation from the impression and management recommendations stated within the original breast imaging reports. A bilingual, fellowship-trained academic breast radiologist (not an author of this study) with 16 years of clinical radiology experience, who is certified in both English and Spanish by LanguageLine Solutions, an on-demand interpretation and translation service (www.LanguageLine.com), reviewed all ChatGPT-generated outputs.

Readability Reference Framework

To categorize and interpret readability scores across the three outputs (original, ChatGPT English, and ChatGPT Spanish), we used multiple readability formulas: the Flesch-Kincaid Reading Level (FKRL), Flesch Reading Ease Score (FRES), Fernández-

Huerta, and Szigriszt-Pazos. We developed a unified reference table (Table 1) aligning these formulas with commonly accepted U.S. grade-level equivalents and qualitative readability descriptors (e.g., "Easy," "Difficult") based on prior literature [21-25]. This framework allowed for consistent binning of readability data across English and Spanish texts and facilitated side-by-side comparison.

FRES	FKRL	Fernández-Huerta	Szigriszt-Pazos	Grade Level	Readability Level
91-100	4	91-100	86-100	4	Very Easy
81-90	5	81-90	76-85	5	Easy
71-80	6	71-80	66-75	6	Relatively Easy
61-70	7-8	61-70	51-65	7-8	Standard
51-60	9-10	51-60	36-50	9-10	Relatively Difficult
31-50	11-12	31-50	16-35	11-12	Difficult
0-30	College	0-30	0-15	College	Very Difficult

Table 1: Reference table comparing the Flesch Reading Ease Score (FRES), Flesch-Kincaid Reading Level (FKRL), Fernández-Huerta, and Szigriszt-Pazos indices for cross comparison with commonly accepted U.S. grade-level equivalents and qualitative readability descriptors.

Statistical Analysis

We compared 50 original reports with their corresponding AI-simplified versions across three key characteristics: Word count, Flesch Reading Ease Score (FRES), and Flesch-Kincaid Reading Level (FKRL). For each characteristic, descriptive statistics were calculated, including measures of central tendency (mean, median) and variability (standard deviation, range). The distribution of differences between paired samples was assessed, and Shapiro-Wilk tests were performed to evaluate normality. Paired t-tests were used to test for statistically significant differences when normality assumptions were met, and Wilcoxon signed-rank tests were applied as a non-parametric alternative when normality was violated. This analytical framework for readability comparison was adapted from Li et al.[5]. All analyses were performed using Python (www.python.org).

Results

Table 2 details the study characteristics and readability metrics for the fifty breast imaging reports included in this study.

#	Reddit post/Report					ChatGPT Output: English				ChatGPT Output: Spanish			
	Word Count	FRES	FKRL	Modality	BI-RADS	Accuracy	Word Count	FRES	FKRL	Accuracy	Word Count	Fernández-Huerta	Szigriszt-Pazos
1	476	34.1	12.8	Mg/US	5	Yes	419	67.2	8.4	Yes	434	93	89.7
2	478	29.9	12.9	Mg/US	4	Yes	485	61.4	9.5	Yes	371	88.1	84.7
3	378	43.9	9.5	Mg/US	4	Yes	327	60.6	9.3	Yes	302	89.5	85.9
4	220	29.3	11.9	Scr	0	Yes	288	56.7	9.8	Yes	252	88.8	85.2
5	202	34.1	11.1	Mg/US	4	Yes	272	63.9	8.6	Yes	292	87.6	83.9
6	61	26.3	13.4	Scr	0	Yes	155	55.6	10.3	Yes	119	83.2	79.7
7	157	35.5	11.8	Mg/US	4	Yes	241	54.1	10.7	Yes	187	85.1	81.5
8	103	36.1	11.4	Scr	0	Yes	141	51.4	11.1	Yes	165	80.3	76.7
9	300	27.6	12.6	Scr	0	Yes	292	52	10.8	Yes	288	84	80.7
10	188	42	9.9	Mg/US	0	Yes	310	59.8	9.2	Yes	356	90	86.6

11	128	-9.9	17.5	Mg/US	0	Yes	226	42	13	Yes	229	81	77.6
12	108	49	9.4	Scr	0	Yes	186	57.8	9.4	Yes	215	84.5	80.9
13	130	15.2	13.4	Scr	0	Yes	196	64.1	9.1	Yes	198	82.1	78.5
14	187	52.9	8.6	US	3	Yes	215	62.7	9	Yes	220	91.9	88.5
15	256	32.3	11.5	Scr + US	0	Yes	266	62.8	8.9	Yes	298	95.2	91.8
16	288	38.3	10.8	Mg/US	0	Yes	363	61.3	9	Yes	342	87	83.6
17	147	17	14.5	Scr	0	Yes	348	65.7	8.4	Yes	206	92.3	88.8
18	173	29.1	11.9	Scr	0	Yes	272	58.7	9.9	Yes	271	77.8	74
19	381	34.5	11.8	Mg/US	0	Yes	371	58.3	9.3	Yes	433	84.8	81.2
20	73	23.1	13	Scr	0	Yes	159	55.4	9.9	Yes	166	90.5	87
21	81	31.7	13.9	Mg/US	4	Yes	176	59.2	9.4	Yes	158	89.1	85.7
22	151	43.7	9.2	Mg/US	5	Yes	222	56.1	9.6	Yes	186	87.3	83.7
23	199	43.7	9	Mg/US	3	Yes	190	62.9	9.2	Yes	226	92.5	89.1
24	120	16.7	13.7	Scr	6	Yes	154	48.9	11.2	Yes	189	82.5	78.8
25	128	25.8	12.5	Scr	0	Yes	155	52.8	9.4	Yes	191	91.3	87.8
26	217	39.7	10.7	Scr	0	Yes	251	52.4	10.7	Yes	208	87.5	84
27	127	40.1	10.8	Scr	0	Yes	217	66.1	8.5	Yes	172	85.8	82.7
28	111	14.5	14.3	Scr	0	Yes	157	56.4	9.7	Yes	162	78.8	75.2
29	103	35.5	10.4	US	4	Yes	195	66.7	8	Yes	129	81.1	77.3
30	154	30.4	13.8	Mg/US	4	Yes	205	68.3	7.7	Yes	188	90.3	87
31	330	31.6	11.7	Mg/US	3	Yes	260	59.2	9.6	Yes	355	87.2	83.7
32	105	56.1	9.8	Mg/US	3	Yes	227	64.4	9.4	Yes	248	91.2	88.1
33	719	54.6	8.5	Bx + Mg/US	5	Yes	302	53.8	10.2	Yes	260	91.1	87.7
34	241	42	9.6	Mg/US	3	Yes	265	61.6	8.8	Yes	276	89.2	85.8
35	166	30	11.8	Scr	0	Yes	228	59.2	10.6	Yes	196	79.7	76
36	163	20.6	14	Scr	0	Yes	222	57.3	10.7	Yes	236	82.3	78.6
37	190	41.7	11.3	Mg/US	5	Yes	246	65.9	8.4	Yes	213	88.2	84.8
38	245	39.4	10.9	Mg/US	3	Yes	169	59	9.2	Yes	192	93.2	90
39	472	43.9	9.6	Mg/US	4	Yes	295	55.7	9.8	Yes	290	82.1	78.5
40	105	23.4	12.9	Dx Mg	4	Yes	157	53.7	11.3	Yes	149	70.9	67.5

41	147	31.6	11.5	Dx Mg	3	Yes	223	65	7.9	Yes	232	83.1	79.4
42	194	37.1	10.8	Scr	0	Yes	279	66.2	9	Yes	220	88.3	84.9
43	306	46.2	10.8	Mg/US	4	Yes	380	61.8	10.5	Yes	353	87.5	84.2
44	243	48.8	9.6	Mg/US	4	Yes	358	65.2	9.4	Yes	368	99.1	96
45	127	32.3	11.3	Dx Mg	0	Yes	170	61.4	10	Yes	189	90.6	87.1
46	78	32.3	11.9	Scr	0	Yes	181	60.7	9.8	Yes	158	87.5	84
47	201	32.5	12.6	Mg/US	3	Yes	224	60.5	10.4	Yes	131	84.4	81.2
48	617	44.7	10	Mg/US	5	Yes	335	61.4	10.2	Yes	315	92.6	89.2
49	297	42.9	12.2	Mg/US	4	Yes	311	62	9.4	Yes	259	88.7	85.4
50	81	23.2	12.5	Scr	1	Yes	110	68.3	6.8	Yes	132	89.3	85.6
Mean	217.0	33.9	11.6				247.9	59.7	9.6		238.5	87.0	83.5
SD	140.4	11.8	1.8				79.6	5.4	1.1		78.6	5.1	5.2
Median	180	34.1	11.6				227.5	60.6	9.4		220	87.6	84.1
Range: Min	61	-9.9	8.5				110	42	6.8		119	70.9	67.5
Range: Max	719	56.1	17.5				485	68.3	13		434	99.1	96

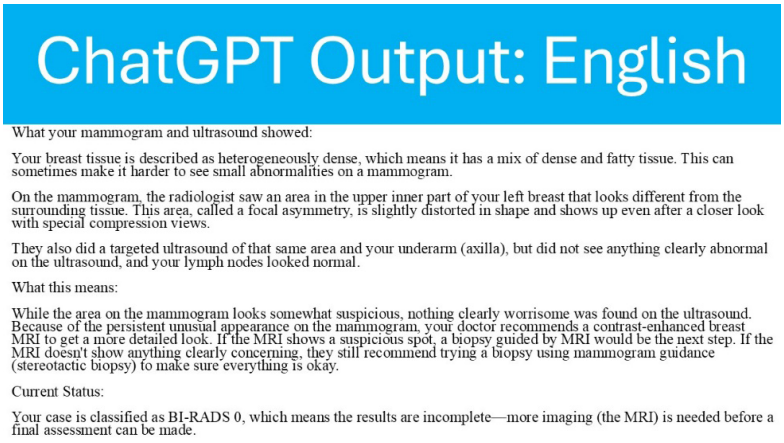
Note: **Bx + Mg/US** = diagnostic mammogram with ultrasound and biopsy combined report, **Dx Mg** = diagnostic mammogram, **FKRL** = Flesch-Kincaid Reading Level, **FRES** = Flesch Reading Ease Score, **Mg/US** = diagnostic mammogram and ultrasound combined report, **Scr** = screening mammogram, **Scr + US** = screening mammogram and ultrasound combined report, **SD** = standard deviation, **US** = ultrasound.

Table 2: Readability data for each original post from Reddit and ChatGPT outputs in English and Spanish.

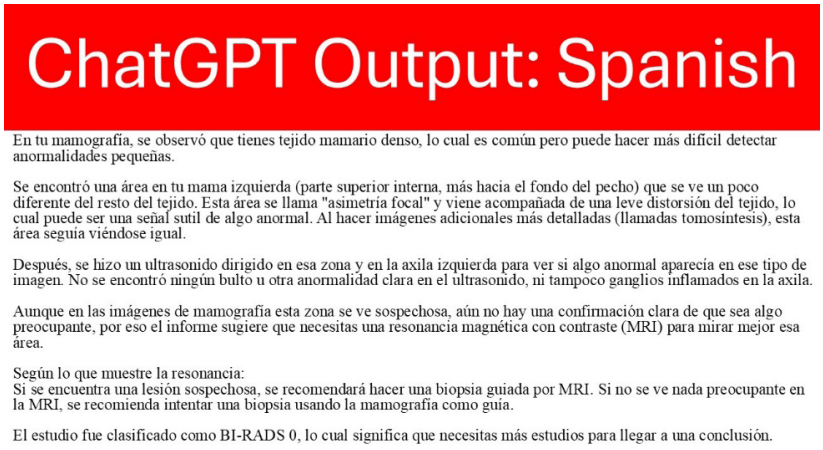
Readability Improvement and Fidelity

ChatGPT English outputs (Figure 1) were significantly easier to read than the original report text from Reddit. Mean Flesch Reading Ease Score (FRES) increased from 33.9 ± 11.8 to 59.7 ± 5.4 ($p < .001$), and mean Flesch-Kincaid Reading Level (FKRL) decreased from 11.6 ± 1.8 to 9.6 ± 1.1 ($p < .001$) (Table 3). Word count was slightly higher in the ChatGPT English outputs compared to the original reports (248 ± 80 vs. 240 ± 140 words), and this difference was statistically significant ($p = .03$), suggesting that readability gains were achieved through simplification of language rather than content reduction. Clinical accuracy, as assessed by the bilingual breast radiologist, was preserved in all 50/50 (100%) English outputs. ChatGPT's Spanish outputs were similarly effective. Fernández-Huerta scores averaged 87.0 ± 5.1 and Szigriszt-Pazos scores averaged 83.5 ± 5.2, both falling in the "easy" readability range. Mean word count (239 ± 79) was again comparable to English outputs ($p = .15$). All 50 Spanish outputs were deemed 100% accurate.

(1A)



(1B)



(1C)

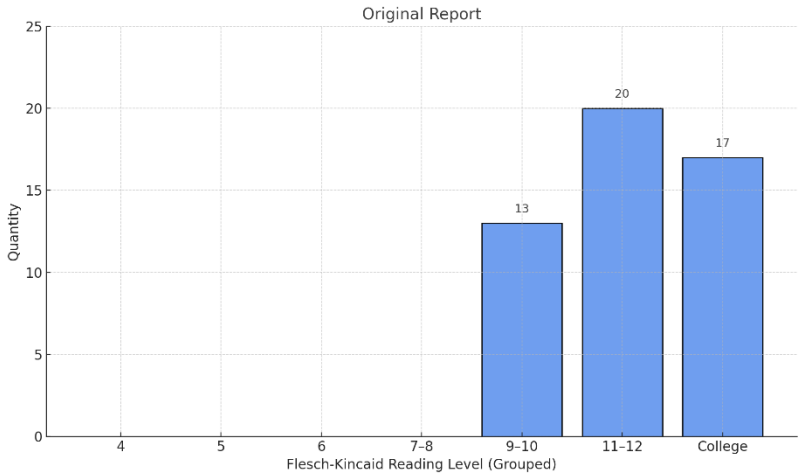
Figure 1: Sample breast imaging report from Reddit **(A)** with correspoding ChatGPT outputs in English **(B)** and Spanish **(C)**.

	Word Count (Mean ± SD)	**FRES (Mean ± SD)**	**FKRL (Mean ± SD)**	**Accuracy**
Report Text from Reddit	240 ± 140	33.9 ± 11.8	11.6 ± 1.8	
ChatGPT Output: English	248 ± 80	59.7 ± 5.4	9.6 ± 1.1	50/50 (100%)
P value	.03	<.001	<.001	
	Word Count (Mean ± SD)	Fernández-Huerta (Mean ± SD)	Szigriszt-Pazos (Mean ± SD)	Accuracy
Report Text from Reddit	240 ± 140			
ChatGPT Output: Spanish	239 ± 79	87.0 ± 5.1	83.5 ± 5.2	50/50 (100%)
P value	.15			
Note: FKRL = Flesch-Kincaid Reading Level, FRES = Flesch Reading Ease Score, SD = standard deviation.				

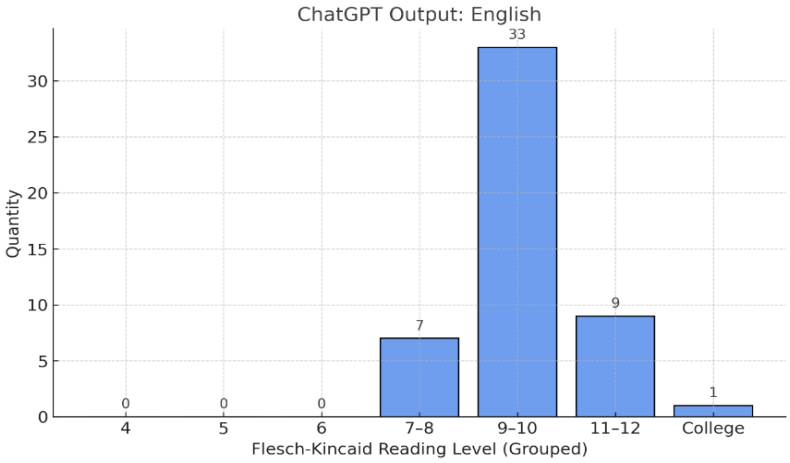
Table 3: Word counts, Flesch Reading Ease Score (FRES), Flesch-Kincaid Reading Level (FKRL), Fernández-Huerta, and Szigriszt-Pazos indices for the original report post from Reddit, English ChatGPT output, and the Spanish ChatGPT output.

Distribution of Reading Level Categories

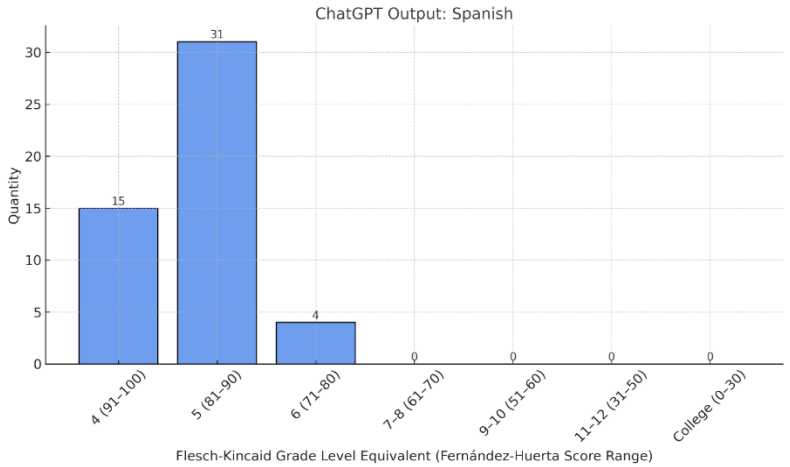
Histogram visualizations (Figure 2) further highlighted these readability improvements. For FKRL, the distribution of grade levels in the original reports skewed toward 11th-12th grade and college-level reading, with 70% of samples above the 9th-grade threshold. In contrast, the ChatGPT English outputs showed a clear shift leftward, with the majority of reports clustering in the 7th-10th grade range, closer to recommended public health standards. For the Spanish outputs, Fernández-Huerta scores were converted to FKRL-equivalent bins for comparison using validated thresholds per Table 1. The Spanish translations overwhelmingly fell into categories equivalent to 4th–6th grade FKRL scores, with no samples exceeding the "7-8" grade band. This shift not only surpassed the English readability gains but also aligned with NIH and AMA recommendations for patient education materials.



(2A)



(2B)



(2C)

Figure 2: Histogram of the reading level of the original report posted on Reddit **(A)**, ChatGPT-simplified English Outputs **(B),** and ChatGPT-simplified Spanish Outputs **(C).**

Subgroup Analysis by Imaging Modality

We stratified reports by imaging modality and procedure complexity to explore whether certain report types were more difficult to simplify (Table 4). Reports involving diagnostic mammography and ultrasound (n=24) were the longest (271 ± 137 words) and had higher FKRL (11.3 ± 2.0) and lower FRES (37.6 ± 12.0) compared to simpler screening mammograms (n=19; 142 ± 60 words; FKRL = 12.3 ± 1.4; FRES = 28.0 ± 9.4). This suggests that complexity and wordiness correlated with lower readability. However, ChatGPT effectively simplified even the longest, most technical reports across all subgroups.

	N	Overall BI-RADS Category (N)	Mean Word Count (SD)	Mean FRES (SD)	Mean FKRL (SD)
Screening Mammogram	19	0 (18) 1 (1)	142 (60)	28.0 (9.4)	12.3 (1.4)
Screening Mammogram and Ultrasound	1	0 (1)	256	32.3	11.5
Diagnostic Mammogram	3	0 (1) 3 (1) 4 (1)	126 (21)	29.1 (4.9)	11.9 (0.9)
Breast Ultrasound	2	3 (1) 4 (1)	145 (59)	44.2 (12.3)	9.5 (1.3)
Diagnostic Mammogram and Breast Ultrasound with Biopsy	1	5 (1)	719	54.6	8.5
Diagnostic Mammogram and Breast Ultrasound	24	0 (4) 3 (6) 4 (10) 5 (4)	271 (137)	37.6 (12.0)	11.3 (2.0)
Note: FKRL = Flesch-Kincaid Reading Level, FRES = Flesch Reading Ease Score, SD = standard deviation.					

Table 4: Readability indices of the original reports posted on Reddit by imaging modality.

Discussion

Collectively, these findings demonstrate that ChatGPT significantly improves the readability of breast imaging reports in both English and Spanish without sacrificing clinical accuracy. Improvements were consistent across all assessed readability metrics and across a diverse set of report types, including screening mammograms, diagnostic studies, ultrasounds, and reports involving biopsy. In English, ChatGPT reduced the reading level by approximately two grade levels on average (from FKRL 11.6 to 9.6), shifting many reports from a "*difficult*" readability range into a more "*standard*" or "*relatively difficult*" range, closer to public health recommendations for patient materials. The distribution histogram confirms that these shifts were not isolated to a few outliers but rather reflected a widespread pattern across the sample. In Spanish, the effect was even more striking. Based on the Fernández-Huerta and Szigriszt-Pazos scales, the vast majority of translated reports achieved "*very easy*" or "*easy*" readability (scores ≥ 80), with no translations falling into the "*very difficult*" or "*difficult*" ranges. This resulted in a noticeable shift toward lower-grade equivalents compared to the original English reports. Importantly, despite these simplifications, all 50 translated outputs retained full clinical accuracy, as verified by the bilingual breast radiologist.

Together, these findings suggest that ChatGPT can be a powerful tool to bridge the communication gap in breast imaging, particularly for Spanish-speaking populations who may face compounded barriers due to both language and literacy challenges. By significantly lowering readability scores and maintaining medical accuracy, ChatGPT outputs may help patients better understand their imaging results, potentially reducing anxiety, improving informed decision-making, and supporting shared decision-making conversations with healthcare providers [26].

Several recent studies have explored the use of LLMs to improve patient understanding of radiology reports, primarily through automated summarization and rewording strategies. In 2016, Hassanpour and Langlotz pioneered this area by developing natural language processing methods to extract key findings from radiology reports for structured summarization [27]. Furthermore, Van Veen et al. introduced RadAdapt, an LLM pipeline using lightweight domain adaptation to fine-tune models specifically on radiology reports [28]. Their model produced concise impressions that aligned with expert radiologist assessments, demonstrating both fluency and factual coherence in summarization tasks. Building on this trajectory, in 2025, Yang et al. employed reinforcement learning from AI feedback to align LLM-generated chest CT impressions with radiologist-preferred summaries, achieving ~78% agreement and measurable gains in precision, recall, and F1 scores over unaligned models [29].

Recent work has also begun to examine LLM performance in producing patient-directed summaries. In 2024, Jeblick et al. evaluated ChatGPT for rewriting radiology reports into plain language and found substantial improvements in readability while maintaining high factual accuracy—though some minor omissions were observed in less clinically significant details [12]. Kuckelman et al. examined ChatGPT-4 in the context of musculoskeletal

radiology and demonstrated that the model could consistently produce simplified summaries that retained key diagnostic content, indicating its potential utility in patient communication workflows [30]. Similarly, Park et al. developed a generative-AI pipeline specifically tailored for creating patient-centered radiology reports [31]. Their study reported notable improvements in patients' perceived clarity and ease of understanding, without compromising clinical accuracy or diagnostic integrity. These findings collectively suggest that LLMs may offer scalable solutions for improving the accessibility of radiologic information. However, despite these encouraging advances, most prior research has focused exclusively on English-language reports and clinician-facing summarization tasks, rather than addressing the dual challenges of improving lay readability and meeting the needs of multilingual patient populations. Further work here is indicated.

While this study demonstrates the potential of ChatGPT to improve accessibility of breast imaging reports through simplified Spanish translations, several limitations merit discussion. First, there is no single readability metric validated for both English and Spanish text, limiting true statistical comparison of English and Spanish readability indices. Second, the dataset consisted of de-identified mammography and ultrasound reports, which may not represent the full range of breast imaging language encountered in diverse clinical settings. The selection may still be biased toward reports with ambiguous or complex wording, given the nature of patient-generated content, and may not capture the variability present in institutional documentation styles. Third, the readability metrics used, Flesch-Kincaid Reading Level (FKRL) and Fernández-Huerta, offer useful benchmarks but are inherently limited in their ability to assess true patient comprehension. These indices are based on surface-level textual features like sentence and word length, rather than semantic clarity, numeracy, or cultural relevance. Moreover, while both are well-established, they were developed for general prose and may not fully capture the complexities of bilingual or health-specific communication. Fourth, the study assessed translation quality and clinical accuracy based on expert review by a single bilingual, fellowship-trained academic breast radiologist. While this ensured a high level of clinical consistency, it limits the generalizability of our findings. Broader validation by panels including native Spanish-speaking patients, patient advocates, and multidisciplinary clinicians will be important to assess real-world utility and trust. Fifth, we evaluated only one LLM, ChatGPT-4o, at a single time point. As LLMs continue to evolve rapidly, future research should compare performance across different models (e.g., Gemini, Claude, Copilot) and evaluate the impact of newer model versions, system prompts, and fine-tuning strategies tailored to radiology or patient education.

Conclusion

This study demonstrates that ChatGPT can effectively simplify breast imaging reports in both English and Spanish, substantially improving readability while preserving clinical accuracy. By reducing the average reading level by approximately two grades in English and achieving "easy" or "very easy" readability in Spanish, ChatGPT addresses a critical communication gap for patients with varying health literacy and language needs. These findings highlight the potential of LLMs to enhance patient engagement and comprehension, particularly in breast imaging, where anxiety and misunderstanding of reports are common. As generative AI technologies continue to advance, integrating them into patient-centered communication strategies, while ensuring expert oversight, may help bridge linguistic and literacy barriers, ultimately improving informed decision-making and trust in radiologic care.

Acknowledgements

None.

Ethical Considerations

This study was deemed exempt from IRB review by the IRB. Informed consent was waived because this study was determined not to be research involving human subjects by the IRB.

Funding

None.

Data Availability

The original contributions presented in this study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Alarifi M, Patrick T, Jabour A, Wu M, Luo J. (2021). Understanding patient needs and gaps in radiology reports through online discussion forum analysis. Insights Imaging. 12: 50.
2. Brett J, Bankhead C, Henderson B, Watson E, Austoker J. (2005). The psychological impact of mammographic screening. A systematic review. Psychooncology. 14: 917-38.
3. Pollock JR, Petty SAB, Schmitz JJ, Varner J, Metcalfe AM. (2024). Patient Access of Their Radiology Reports Before and After Implementation of 21st Century Cures Act Information-Blocking Provisions at a Large Multicampus Health System. AJR Am J Roentgenol. 222: e2330343.
4. McKee J. (2025). Radiology Adjusts to The 21st Century Cures Act. Accessed in Online.
5. Li H, Moon JT, Iyer D. (2023). Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. Clin Imaging. 101: 137-141.
6. Rogers C, Willis S, Gillard S, Chudleigh J. (2023). Patient experience of imaging reports: A systematic literature review. Ultrasound. 31: 164-

175.

7. Munawar K, Prabhu V. (2021). Radiology on Reddit: A Content Analysis and Opportunity for Radiologist Engagement and Education. Curr Probl Diagn Radiol. 50: 362-368.

8. Dietrich S, Hernandez E. (2019). Nearly 68 Million People Spoke a Language Other Than English at Home in 2019. Accessed in Online.

9. Villa Camacho JC, Pena MA, Flores EJ. (2021). Addressing Linguistic Barriers to Care: Evaluation of Breast Cancer Online Patient Educational Materials for Spanish-Speaking Patients. J Am Coll Radiol. 18: 919-926.

10. Choudhery S, Xi Y, Chen H. (2020). Readability and Quality of Online Patient Education Material on Websites of Breast Imaging Centers. J Am Coll Radiol. 17: 1245-1251.

11. AlKhalili R, Shukla PA, Patel RH, Sanghvi S, Hubbi B. (2015). Readability assessment of internet-based patient education materials related to mammography for breast cancer screening. Acad Radiol. 22: 290-295.

12. Jeblick K, Schachtner B, Dexl J. (2024). ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 34: 2817-2825.

13. Tepe M, Emekli E. (2024). Assessing the Responses of Large Language Models (ChatGPT-4, Gemini, and Microsoft Copilot) to Frequently Asked Questions in Breast Imaging: A Study on Readability and Accuracy. Cureus. 16: e59960.

14. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM. (2023). Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. Digit Health. 9: 20552076231221620.

15. Gibson D, Jackson S, Shanmugasundaram R. (2024). Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment. J Med Internet Res. 26: e55939.

16. Eppler MB, Ganjavi C, Knudsen JE. (2023). Bridging the Gap Between Urological Research and Patient Understanding: The Role of Large Language Models in Automated Generation of Layperson's Summaries. Urol Pract. 10: 436-443.

17. Butler JJ, Harrington MC, Tong Y. (2024). From jargon to clarity: Improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. Foot Ankle Surg. 30: 331-337.

18. Temperley HC, O'Sullivan NJ, Mac Curtain BM. (2024). Current applications and future potential of ChatGPT in radiology: A systematic review. J Med Imaging Radiat Oncol. 68: 257-264.

19. Sarangi PK, Lumbani A, Swarup MS. (2023). Assessing ChatGPT's Proficiency in Simplifying Radiological Reports for Healthcare Professionals and Patients. Cureus. 15: e50881.

20. Rogasch JMM, Metzger G, Preisler M. (2023). ChatGPT: Can You Prepare My Patients for [J Nucl Med. 64: 1876-1879.

21. Cherla DV, Sanghvi S, Choudhry OJ, Liu JK, Eloy JA. (2012). Readability assessment of Internet-based patient education materials related to endoscopic sinus surgery. Laryngoscope. 122: 1649-1654.

22. Fajardo-Delgado D, Rodriguez-Coayahuitl L, Sánchez-Cervantes MG. (2025). Readability Formulas for Elementary School Texts in Mexican Spanish Applied Sciences. 15: 7259.

23. O'Sullivan L, Sukumar P, Crowley R, McAuliffe E, Doran P. (2020). Readability and understandability of clinical research patient information leaflets and consent forms in Ireland and the UK: a retrospective quantitative analysis. BMJ Open. 10: e037994.

24. Nash E, Bickerstaff M, Chetwynd AJ, Hawcutt DB, Oni L. (2023). The readability of parent information leaflets in paediatric studies. Pediatr Res. 94: 1166-1171.

25. Graham C, Reynard JM, Turney BW. (2015). Consent information leaflets - readable or unreadable? J Clin Urol. 8: 177-182.

26. van der Mee FAM, Ottenheijm RPG, Gentry EGS. (2025). The impact of different radiology report formats on patient information processing: a systematic review. Eur Radiol. 35: 2644-2657.

27. Hassanpour S, Langlotz CP. (2016). Information extraction from multi-institutional radiology reports. Artif Intell Med. 66: 29-39.

28. Van VD, Van UC, Attias M. (2023). RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models. presented at: In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; Toronto, Canada.

29. Yang L, Zhou Y, Qi J. (2025). Aligning large language models with radiologists by reinforcement learning from AI feedback for chest CT reports. Eur J Radiol. 184: 111984.

30. Kuckelman IJ, Wetley K, Yi PH, Ross AB. (2024). Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. Skeletal Radiol. 53: 1621-1624.

31. Park J, Oh K, Han K, Lee YH. (2024). Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. Sci Rep. 14: 13218.