Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264309

BREAST IMAGING

ORIGINAL ARTICLE

©Copyright 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Prompt-dependent readability and accuracy of Generative Pre-trained Transformer-5 thinking responses to breast imaging questions derived from European Society of Breast Imaging patient recommendations

Ebru Düşünceli Atman[1]
Sena Bozer Uludağ[1]
Çağlar Uzun[1]
Zeynep Eskalen[1]
Gökhan Uludağ[2]

[1]Ankara University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

[2]University of Health Sciences Türkiye, Ankara Bilkent City Hospital, Clinic of General Surgery, Division of Gastroenterology, Ankara, Türkiye

PURPOSE

Although professional societies provide standardized resources, the reliability of online health content remains inconsistent, particularly as large language models (LLMs) become a growing source of patient information. Therefore, the primary objective of this study was to determine whether patient-focused prompt optimization improves the readability of Generative Pre-trained Transformer-5 (GPT-5) Thinking responses compared with standard prompting while evaluating potential changes in expert-rated factual accuracy using educational materials published by the European Society of Breast Imaging (EUSOBI).

METHODS

All GPT-5 Thinking outputs were prospectively generated in January 2026. The dataset included frequently asked questions (FAQs) and special information notes (SINs) derived from EUSOBI articles, each submitted under two distinct prompts (standard-Prompt 1 vs. patient-focused-Prompt 2). Readability was quantified using four indices with an averaged composite "reading grade level." Accuracy was scored by two radiologists using a 5-point Likert scale.

RESULTS

A total of 48 questions (15 FAQs and 33 SIN-derived questions) were each answered under two prompt conditions, yielding 96 responses for analysis. Prompt 2 produced significantly lower readability levels than Prompt 1 across all indices ($P < 0.001$). These improvements occurred without significant differences in word count ($P = 0.12$) or sentence count ($P = 0.63$). The mean accuracy scores remained high and comparable between prompts for all questions (4.54 ± 0.63 vs. 4.52 ± 0.59, $P = 0.75$). The inter-rater agreement was almost perfect for both the FAQ and SIN-derived question sets ($\kappa = 0.82$ and 0.81, respectively).

CONCLUSION

Patient-focused prompting substantially reduced the linguistic complexity and reading grade level of GPT-5 Thinking responses without a statistically significant reduction in expert-rated accuracy, demonstrating its ability to produce more accessible breast imaging content, although the resulting grade level remained above widely recommended patient-education thresholds.

CLINICAL SIGNIFICANCE

A simple, low-cost prompt modification can substantially improve the readability of LLM-generated patient education content in breast imaging without a measurable reduction in expert-rated accuracy, although further work is needed to confirm this translates into improved patient comprehension.

KEYWORDS

Women's health, breast imaging, artificial intelligence, large language model, ChatGPT, readability, patient education; health literacy

Handling editor: Görkem Durak

Corresponding author: Ebru Düşünceli Atman

E-mail: ebrumd2001@yahoo.com

Received 25 July 2026; revision requested 10 August 2026; last revision requested 04 October 2026; accepted 16 October 2026.

Epub: 12.11.2026

DOI: 10.4274/dir.2026.264309

You may cite this article as: Düşünceli Atman E, Bozer Uludağ S, Uzun Ç, Eskalen Z, Uludağ G. Prompt-dependent readability and accuracy of Generative Pre-trained Transformer-5 thinking responses to breast imaging questions derived from European Society of Breast Imaging patient recommendations. *Diagn Interv Radiol*. 11 November 2026 DOI: 10.4274/dir.2026.264309 [Epub Ahead of Print].

Breast imaging plays a crucial role in the comprehensive management of breast diseases, from population-based screening of asymptomatic women to the diagnostic evaluation of clinical or imaging abnormalities, particularly breast cancer. The primary imaging modalities, mammography, ultrasound (US), and magnetic resonance imaging (MRI), offer essential guidance for subsequent treatment planning and follow-up care. Each of these techniques addresses specific clinical questions, and understanding the indications, benefits, and risks associated with each examination is essential for patients. In an environment characterized by the rapid, unfiltered spread of online information, providing clear, reliable, evidence-based explanations to patients is more vital than ever. Studies have confirmed that a patient's health literacy is a major determinant of their ability to navigate breast imaging pathways and is an independent predictor of participation in organized breast cancer screening programs.[1] Therefore, strategies dedicated to improving health literacy may increase adherence to screening recommendations by promoting a deeper understanding of potential outcomes and risks.

To promote consistent and high-quality patient education, the European Society of Breast Imaging (EUSOBI) published "Recommendations for Women's Information," which provides clear, consensus-based explanations aligned with guidelines and comprises three separate articles focusing on mammography, MRI, and US, respectively.[2-5] Each article contains five frequently asked questions (FAQs) accompanied by official society-provided responses, along with several embedded special information notes (SINs) for detailed patient instruction.

Main points

- Breast cancer is the most common type of cancer among women, and women of all ages need to have easy access to accurate information regarding early diagnosis and screening.
- In this context, web-based search engines, social media, and artificial intelligence models provide women of all educational levels with information that may be accurate or inaccurate, sufficient or insufficient.
- A patient-focused prompt substantially improved the readability of Generative Pre-trained Transformer-5 Thinking's breast imaging responses relative to a standard prompt, without a statistically significant reduction in expert-rated clinical accuracy.

Healthcare communication is undergoing a significant transformation, driven primarily by the advancements in artificial intelligence (AI). Large language models (LLMs), one of the most promising developments in AI, have reshaped natural language processing. Recently, LLMs have become widely accessible by demonstrating powerful capabilities for generating clinically relevant health content and providing appropriate recommendations.[6-12] LLMs offer an effective way to address common patient inquiries by using extensive training data to deliver clear, guideline-aligned explanations. These systems thereby strengthen patient health literacy while simultaneously reducing the substantial informational workload carried by healthcare professionals.[13-15]

OpenAI's Generative Pre-trained Transformer (GPT) series has advanced rapidly, enabling human-like dialogue through deep contextual learning. From GPT-3.5 to multimodal GPT-4, successive architectural improvements and reinforcement learning with human feedback have increased the accuracy, factual alignment, and coherence of generated text. The ChatGPT platform further popularized this technology by offering conversational access for diverse applications in education, research, and healthcare communication.[16] Building on these iterations, GPT-5 Thinking represents the newest generation of this model family. According to OpenAI's own product announcement, GPT-5 Thinking is described as a faster and more capable, unified model that reduces hallucinations, more reliably follows instructions, and shows its best performance to date on health-related benchmarks;[17] these vendor-reported claims have not yet been independently validated in peer-reviewed literature, which is part of the motivation for the present study.

Given these advancements and their relevance to patient communication, this study evaluates the prompt-dependent readability and expert-rated accuracy of GPT-5 Thinking, one of the latest reasoning-focused model versions available at the time of study execution, using EUSOBI's published patient recommendations [mammography (2012), MRI (2015), and US (2018)] as a standardized clinical benchmark. We hypothesized that a patient-focused prompt would significantly reduce the composite reading grade level of GPT-5 Thinking's outputs toward the recommended threshold for patient education materials, without causing a clinically meaningful reduction in accuracy.

Methods

Study design

Because this study used prospectively generated LLM data and did not involve identifiable patient data or direct human participant research, ethics committee approval and/or written informed consent were not required.

All outputs were generated with the ChatGPT (GPT-5 Thinking; OpenAI) web interface in January 2026. Initially, the five FAQs from each modality-specific article (Electronic Supplementary Table 1) were individually submitted to GPT-5 Thinking on the same day, and the generated answers were documented. No additional prompting or input manipulation was applied to influence or optimize the responses of the chatbot. Each query was entered in a new conversation window to minimize potential bias from previous interactions, ensuring that earlier inputs did not affect subsequent responses. Separately, the SINs from each article were converted into appropriate open- or closed-ended questions with the assistance of GPT-5 Thinking. This task was conducted via a distinct account to ensure impartiality and prevent influence from prior sessions. The model was instructed to transform the original text paragraphs into questions without further contextual guidance, ensuring the generated questions reflected independent outputs. The prompt was as follows: "Convert the following paragraph into an [open-ended/closed-ended] question in English, preserving the original meaning without adding or omitting any information." To prevent investigator bias and maintain strict input standardization, SIN-derived questions were generated automatically by GPT-5 Thinking without manual rephrasing or editorial filtering prior to answer generation. The 15 verbatim EUSOBI FAQs represented expert-authored, human-validated baseline queries, whereas the 33 SIN items evaluated the model's unassisted capacity to convert narrative clinical text into structured questions. All eligible summary statements and recommendation items across the three selected EUSOBI patient guidance documents were included in the analysis without any exclusions. Electronic Supplementary Table 2 presents the questions generated from the SINs sections of each EUSOBI article.

Two prompts, standard and patient-focused, were used to generate and compare responses for subsequent readability and

Table 1. Likert scale to assess the accuracy of Generative Pre-trained Transformer-5 Thinking responses

Score	Description
1–Completely incorrect	Contains major factual errors or conflicts with guidelines; may include unsafe or contraindicated advice; key points are missing or wrong.
2–Mostly incorrect	Multiple inaccuracies or large gaps; unclear or misleading statements; weak guideline concordance; not suitable without substantial revision.
3–Moderately correct	Core idea is partly correct but important details are missing or unclear; minor errors or outdated points may appear; limited actionability.
4–Mostly correct	Factually sound and guideline-aligned with small omissions or wording issues; provides generally actionable, patient-appropriate guidance.
5–Completely correct	Accurate, clear, and safe; covers essential points with adequate detail; patient-friendly language with concrete, appropriate next steps.

accuracy scoring: Prompt 1: "Answer the following question in English"; Prompt 2: "For a patient with no medical background, answer the following question in English." For each EUSOBI article, one question was submitted once per prompt, yielding two responses per question. The patient-focused prompt wording was adapted from persona-priming strategies that evaluate LLM responses for lay audiences. Alternative patient-focused phrasings were not compared. The magnitude of the readability effect reported here should therefore be interpreted as specific to this prompt formulation rather than as an upper bound on achievable simplification. To ensure comparable readability and accuracy assessments, model outputs were standardized prior to scoring by removing non-prose formatting, consistent with prior studies.[18-20] Non-contentual ancillary segments, including salutations and conversational scaffolding, identity/capability or legal disclaimers, user-interface prompts, and links or footer notes unrelated to the clinical explanation, were also removed. Text normalization removed an average of 23.6 ± 7.2 words (12.1%) in Prompt 1 and 20.4 ± 6.1 words (11.3%) in Prompt 2, with no significant difference between conditions. Normalization was performed by one operator who was blinded to the explicit prompt labels and did not participate in the qualitative evaluations, following the predefined rule set described above. In the final step, remaining content was reconstructed into continuous plain-prose sentences with standard sentence boundaries (periods, capitalization), without adding, deleting, or rephrasing any clinical words. The same protocol was applied to both prompts; no lexical simplification was applied. Detailed normalization rules are provided in Electronic Supplementary Table 3. All normalized GPT-5 Thinking responses used for readability and accuracy assessment are provided in Electronic Supplementary Table 4. Because the standard web interface does not expose generation parameters (e.g., temperature or top-*P*), these values could not be controlled or recorded. Additionally, specific underlying model snapshot builds or system-level prompt layers active at the time of generation were not retrievable. To maximize reporting transparency and ensure auditability, all raw GPT-5 Thinking responses and their corresponding normalized texts were also archived, and a PDF screenshot was saved for each prompt under each question.

The primary endpoint was the difference in readability (composite reading grade level) between standard (Prompt 1) and patient-focused (Prompt 2) responses. Secondary endpoints were the difference in expert-rated accuracy between the two prompts, and the absolute readability and accuracy levels achieved under each prompt individually. Official EUSOBI patient information materials were selected to serve as an authoritative, standardized clinical knowledge base for query formulation. The study design specifically evaluated the comparative impact of prompt modification (Prompt 1 vs. Prompt 2) on LLM output readability and accuracy, rather than assessing or auditing the baseline readability of the underlying society-authored source texts.

Readability assessment

The readability of GPT-5 Thinking responses was quantified using four established indexes, which correspond to a reading grade: Flesch–Kincaid Grade Level (FKG), Gunning Fog (FOG), Automated Readability Index (ARI), and Coleman–Liau Index (CLI). The scores correspond to educational levels: a grade of 6 indicates readability for sixth-grade, 12 approximates a high-school senior level, and 17 reflects college-graduate readability.[21-23] Consistent with previous readability studies, the four readability indexes were computed for each response and averaged to create a composite score, which is referred to as the "reading grade level (RGL)".[18,19,24,25] Readability metrics were calculated using the quanteda.textstats package in R (R version 4.3.2). Length metrics, including word and sentence counts, were also recorded to support index calculations.

Accuracy assessment

The scientific accuracy and clinical appropriateness of the GPT-5 Thinking responses were independently evaluated by two radiologists with 3 and 22 years of experience in breast imaging, respectively. The responses were de-identified and randomized, and the reviewers were blinded to the explicit prompt labels. Each answer was scored on a 5-point Likert scale (1: completely incorrect, 5: completely correct) (Table 1). The inter-rater agreement was quantified using quadratic-weighted Cohen's kappa (κ) with 95% confidence intervals (CIs). For responses on which the two radiologists' independent ratings matched, that shared score was retained as the final accuracy score. For responses with an initial scoring discrepancy, the final accuracy score for each response was defined as the arithmetic mean of the two independent ratings. Radiologists rated each response for clinical correctness, medical appropriateness, and patient safety, using the EUSOBI recommendations as the primary knowledge base and applying their own clinical expertise; responses that contradicted EUSOBI guidance were scored lower. Scoring was not performed against a formal item-by-item checklist.

Figure 1 presents a schematic overview of the study design, which includes study materials, question sets, LLM generation and prompts, endpoints (readability and accuracy), and statistical analysis.

Statistical analysis

Statistical analyses were performed using SPSS for Windows, version 25.0 (IBM Corp., Armonk, NY, USA). The Shapiro–Wilk test was used to assess normality. Descriptive

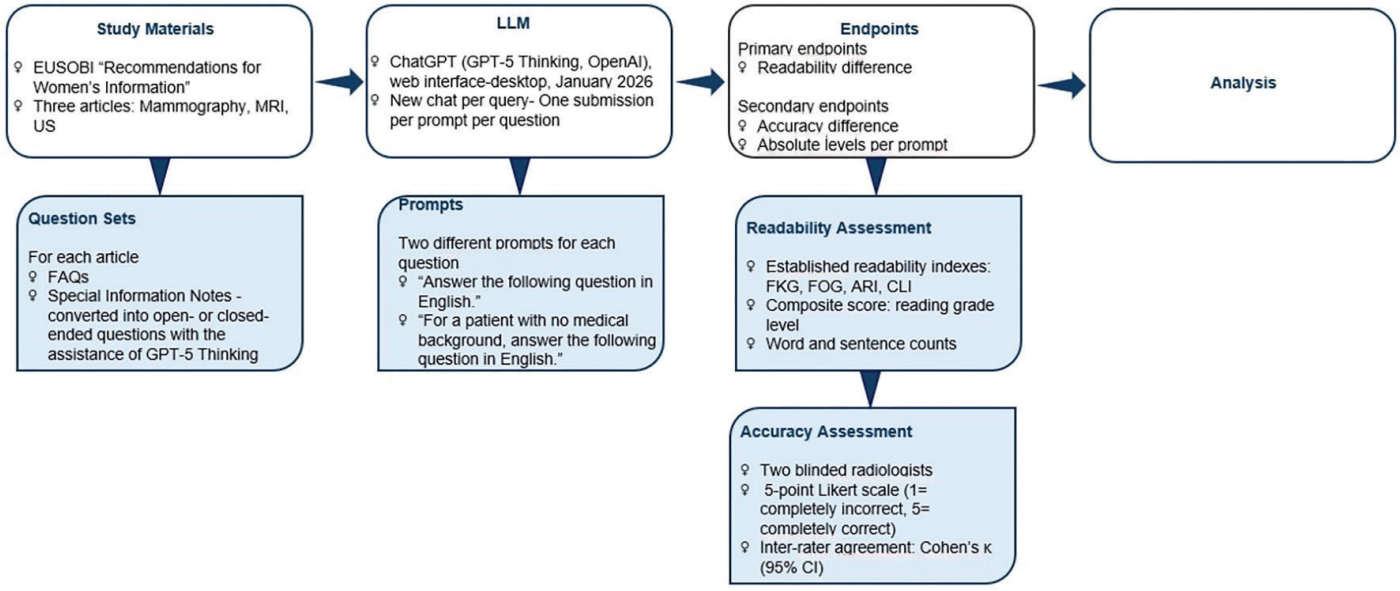


Figure 1. Flowchart of the study design. ARI, Automated Readability Index; CI, confidence interval; CLI, Coleman–Liau Index; EUSOBI, European Society of Breast Imaging; FAQs, frequently asked questions; FKG, Flesch–Kincaid Grade Level; FOG, Gunning Fog; GPT, Generative Pre-trained Transformer; LLM, large language model.

statistics are presented as mean ± standard deviation (SD) for normally distributed variables. To provide a comprehensive overview, non-normally distributed variables are reported as mean ± SD alongside median (interquartile range). Descriptive readability and accuracy statistics were evaluated at the response level (n = 96 total outputs), whereas paired comparisons between prompt conditions (Prompt 1 vs. Prompt 2) operated at the question level (n = 48 pairs). For paired comparisons between prompts for the same question, paired t-tests were used to evaluate differences if normality was achieved, and Cohen's d was reported for paired samples, along with 95% CIs for the mean difference. If normality was not satisfied, the Wilcoxon signed-rank test was employed, with the rank-biserial r reported. Inter-rater agreement for the 5-point ratings was assessed using quadratic-weighted Cohen's κ, along with 95% CIs. All statistical tests were two-sided, with a significance level set at $\alpha = 0.05$.

Results

A total of 48 items were analyzed, corresponding to 15 FAQs and 33 SINs-derived questions. Each item produced two GPT-5 Thinking responses—one per prompt—yielding 96 total outputs for evaluating readability and accuracy. The 15 FAQs comprised 5 questions for each breast imaging modality, whereas the SINs-derived questions included 9 for mammography, 17 for MRI, and 7 for US.

Readability assessment

Across both the 15 FAQs and the 33 SINs-derived questions, Prompt 2 consistently yielded lower readability grade levels than Prompt 1 across all four indices, indicating easier readability. For the FAQ set, the average reductions were 2.78 points for FKG, 2.90 for FOG, 3.26 for ARI, and 3.42 for CLI, with large effect sizes (all $P < 0.001$, paired t-test; Cohen's d range, −1.56 to −1.91); for the SINs-derived questions, the corresponding mean reductions were 2.45, 2.76, 2.90, and 2.68 points, respectively (all $P < 0.001$, paired t-test; Cohen's d range, −1.18 to −1.66). The composite RGL also decreased significantly with Prompt 2 in both datasets, by 3.09 points for FAQs (15.45 ± 1.99 vs. 12.36 ± 2.07; $P < 0.001$; d = −1.73) and by 2.70 points for SINs-derived questions (15.06 ± 2.07 vs. 12.37 ± 1.76; $P < 0.001$; d = −1.73), demonstrating a consistent, substantial improvement in readability across all question types. Similarly, the composite RGL for all questions decreased significantly by 2.82 points with Prompt 2 (15.18 ± 2.03 vs. 12.37 ± 1.84; $P < 0.001$; d = −1.73). Under Prompt 2, 5 of 48 (10.4%) responses achieved a grade level within the widely recommended sixth- to eighth-grade range, whereas none of the responses achieved this threshold under Prompt 1. Length-related metrics, including word and sentence counts, showed no statistically significant differences between prompts in either dataset. Non-parametric tests revealed no meaningful change in the median word or sentence counts for the FAQs, and for the SINs-derived questions, word and sentence counts remained comparable between prompts ($P = 0.82$ and $P = 0.36$, Wilcoxon signed-rank test, respectively). These findings indicate that the observed gains in readability with Prompt 2 were achieved without increasing the overall length or structural complexity of the generated text.

Accuracy assessment

The mean accuracy scores on the 5-point Likert scale were as follows: for the FAQs, scores were 4.47 ± 0.69 for Prompt 1 and 4.33 ± 0.67 for Prompt 2 (95% CI: −0.44 to 0.17; P = 0.36, paired t-test); for the SIN-derived questions, scores were 4.58 ± 0.60 for Prompt 1 and 4.61 ± 0.54 for Prompt 2 (95% CI: −0.12 to 0.18; P = 0.68, paired t-test). Overall, across all 48 questions, mean accuracy was 4.54 ± 0.63 for Prompt 1 vs. 4.52 ± 0.59 for Prompt 2 (mean difference, −0.02; 95% CI: −0.16 to 0.12; P = 0.75, paired t-test; Cohen's d = −0.05). These narrow CIs crossing zero demonstrate no statistically significant difference in factual accuracy between prompt conditions, although formal clinical equivalence was not evaluated against a prespecified margin. No individual rating below 2 was assigned. The only ratings of 2 were given by one rater to FAQ US-5 under both prompts (the other rater scored 3; final score 2.5 in each condition). Item-level and per-rater ratings are provided in Electronic Supplementary Table 5. Inter-rater agreement between the two reviewers was almost

Table 2. Comparison of the readability and accuracy scores for Generative Pre-trained Transformer-5 Thinking responses under two prompt conditions

Prompt metrics are based on questions (n = 48) analyzed across two prompt conditions. As a result, a total of 96 responses were evaluated

	Prompt 1	Prompt 2	Δ (95% CI)	Effect size	*P*
FAQs					
Readability					
FKG	14.86 ± 2.09 14.90 (13.05–15.90)*	12.08 ± 1.60	−2.78 (−3.59, −1.98)	−1.91	< ***0.001***
FOG	16.96 ± 2.33	14.06 ± 2.20	−2.90 (−3.78, −2.01)	−1.81	< ***0.001***
ARI	15.04 ± 2.35	11.78 ± 1.79 12.31 (11.27–12.81)*	−3.26 (−4.29, −2.24)	−1.76	< ***0.001***
CLI	14.94 ± 2.40	11.52 ± 1.62	−3.42 (−4.63, −2.20)	−1.56	< ***0.001***
Reading grade level	15.45 ± 1.99	12.36 ± 2.07	−3.09 (−4.08, −2.10)	−1.73	< ***0.001***
Word count	169.2 ± 110.1 145.0 (129.0–170.0) *	131.9 ± 29.9	−37.30 (−95.47, 20.81)	−0.47	*0.12*
Sentence count	11.3 ± 17.2 6.0 (5.5–9.0)*	6.8 ± 2.5	−4.50 (−14.09, 5.16)	−0.43	*0.69*
Accuracy**	4.47 ± 0.69 5.0 (4.0–5.0)*	4.33 ± 0.67 4.5 (4.0–5.0)*	−0.13 (−0.44, 0.17)	−0.24	*0.36*
SINs-derived questions					
Readability					
FKG	14.75 ± 1.84	12.30 ± 1.21	−2.45 (−2.98, −1.93)	−1.66	< ***0.001***
FOG	16.50 ± 2.22	13.74 ± 1.61	−2.76 (−3.37, −2.14)	−1.60	< ***0.001***
ARI	14.61 ± 2.37	11.71 ± 1.54	−2.90 (−3.53, −2.27)	−1.63	< ***0.001***
CLI	14.40 ± 2.78	11.72 ± 1.72	−2.68 (−3.48, −1.87)	−1.18	< ***0.001***
Reading grade level	15.06 ± 2.07	12.37 ± 1.76	−2.70 (−3.25, −2.15)	−1.73	< ***0.001***
Word count	173.2 ± 55.7 169.00 (127.00–205.00)*	171.6 ± 53.2	−1.60 (−16.94, 13.79)	−0.04	*0.82*
Sentence count	8.3 ± 2.8	8.7 ± 2.8	0.40 (-0.51, 1.35)	0.16	*0.36*
Accuracy**	4.58 ± 0.60 5.0 (4.0–5.0)*	4.61 ± 0.54 5.0 (4.0–5.0)*	0.03 (−0.12, 0.18)	0.07	*0.68*
All questions					
Readability					
FKG	14.78 ± 1.90	12.23 ± 1.33	−2.55 (−2.98, −2.13)	−1.75	< ***0.001***
FOG	16.64 ± 2.24	13.84 ± 1.80	−2.80 (−3.28, −2.31)	−1.68	< ***0.001***
ARI	14.75 ± 2.35	11.73 ± 1.60	−3.02 (−3.54, −2.50)	−1.68	< ***0.001***
CLI	14.57 ± 2.65	11.66 ± 1.67	−2.91 (−3.56, −2.25)	−1.29	< ***0.001***
Reading grade level	15.18 ± 2.03	12.37 ± 1.84	−2.82 (−3.29, −2.35)	−1.73	< ***0.001***
Word count	172 ± 75.7 151.50 (127.25–197.75)*	159.2 ± 50.4	−12.8 (−33.62, 7.12)	−0.23	*0.12*
Sentence count	9.2 ± 9.8 7.50 (6.00–10.25)*	8.1 ± 2.8 7.00 (6.00–9.25)*	−1.10 (−3.97, 1.77)	−0.07	*0.63*
Accuracy**	4.54 ± 0.63 5.0 (4.0–5.0)*	4.52 ± 0.59 5.0 (4.0–5.0)*	−0.02 (−0.16, 0.12)	−0.05	*0.75*

* The values are presented as both mean ± SD and median (interquartile range), due to the non-normal distributions.

** Reported accuracy values were calculated from the averaged ratings of two blinded radiologists. Identical scores were retained, and discrepant ratings were resolved as the arithmetic mean.

Δ denotes the paired difference (P2–P1); negative values indicate lower grade level/shorter text with Prompt 2. Effect sizes are reported as Cohen's *d* for parametric tests and as rank-biserial correlation (r) for non-parametric tests. Two-sided tests; $\alpha = 0.05$

ARI, Automated Readability Index; CI, confidence interval; CLI, Coleman–Liau Index; FAQs, frequently asked questions; FKG, Flesch–Kincaid Grade Level; FOG, Gunning Fog; SINs, special information notes.

perfect, with quadratic-weighted Cohen's κ exceeding 0.80 across both datasets (0.82, 95% CI: 0.72–0.90 for FAQs and 0.81, 95% CI: 0.71–0.90 for SIN-derived questions).

Table 2 summarizes the comparative results for the readability and accuracy scores for both question sets. Boxplots illustrating the overall shift toward lower grade levels under Prompt 2 are presented in Figure 2.

Discussion

In this exploratory study, we evaluated the performance of GPT-5 Thinking in generating appropriate patient education materials based on EUSOBI's published breast imaging recommendations for women. Our main findings showed that a patient-focused prompt substantially improved readability compared with a standard prompt, without a statistically significant reduction in expert-rated accuracy. However, despite these statistically significant gains, only a small fraction of outputs reached the recommended health literacy grade threshold, with the majority remaining above target comprehension levels. Taken together, our results demonstrate that although GPT-5 Thinking can produce expert-verified factual content for patient-centered breast imaging communication, targeted prompt optimization is essential to bridge the gap between complex radiological concepts and health literacy needs.

Previous studies have shown mixed results regarding which LLM produces the most readable content, indicating that readability performance may vary depending on the task, prompt style, or model version.[13] Although all our model's baseline responses generated under standard prompting remained above the widely recommended sixth- to eighth-grade reading level, the outputs produced using patient-focused prompting achieved consistently lower grade-level scores. This indicates that the structural complexity of model-generated text is highly sensitive to prompt framing, demonstrating that readability can be improved through targeted prompt design. However, despite the statistically significant reduction in reading difficulty, only 10.4% (5/48) of patient-focused responses successfully reached the recommended sixth- to eighth-grade threshold, leaving the vast majority of simplified outputs (89.6%, 43/48) above ideal patient literacy targets. This result aligns with the literature highlighting LLMs' inherent difficulty in achieving linguistic simplicity and optimal health literacy standards without further intervention. In agreement with this finding, in a study evaluating ChatGPT's reliability in answering common imaging-related questions under both unprompted and prompted conditions, Gordon et al. found that the FKG level remained high at 13.6 (95% CI: 12.9–14.2) and was statistically unchanged by the prompt [13.0 (95% CI: 12.4–13.6); $P = 0.2$).[26] Crucially, none of the responses reached the eighth-grade readability level recommended for patient-facing materials. However, when ChatGPT is prompted to simplify the existing content, it demonstrates significant potential. Haver et al.[27] found that ChatGPT significantly improved the mean reading ease (46 vs. 70; $P < 0.001$) and readability (13 vs. 8.9; $P < 0.001$). Furthermore, all simplified responses met the criteria for adequate reading ease, compared with only 2 of the 25 original responses ($P < 0.001$). This success confirms that the model can significantly reduce linguistic complexity, although it requires direct instructional prompts to achieve high readability.

Prior research evaluating LLM-generated medical information has shown that these systems are generally capable of producing accurate responses to patient-facing questions across a range of clinical domains. Several studies examining earlier GPT models have similarly reported high accuracy rates when addressing common patient queries across various medical fields, including breast radiology.[26-29] Our findings align with this broader literature: GPT-5 Thinking produced answers consistently judged scientifically sound, with no reduction in accuracy when a more patient-oriented prompt was used. Recent work by Tepe and Emekli[13] provides additional context for interpreting our findings. In their evaluation of ChatGPT-4, Microsoft Copilot, and Gemini, the authors used a set of patient-generated questions in breast imaging and reported notable performance differences among LLMs. ChatGPT-4 achieved the highest accuracy (Likert score of 4.95 ± 0.22), outperforming both Gemini and Copilot, whereas the latter two systems showed comparable but lower correctness ratings. Their results suggest that accuracy can vary substantially across platforms, even when models are presented with similar patient-facing prompts. Although our study focused exclusively on GPT-5 Thinking and used guideline-based content rather than public query trends, both studies converge

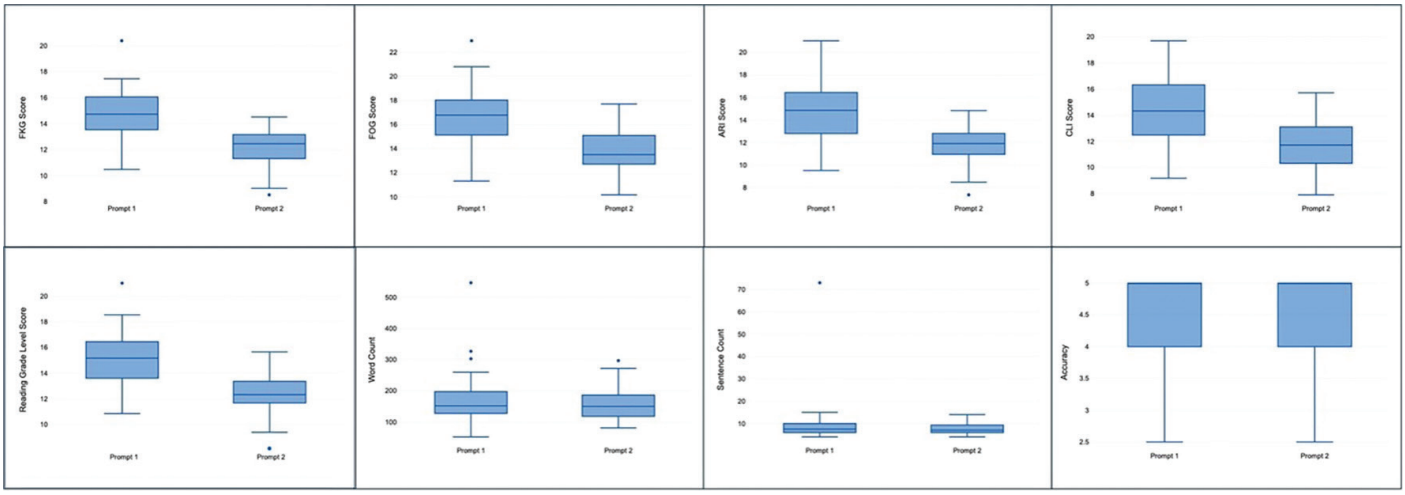


Figure 2. Boxplots showing the readability indices, length metrics, and accuracy scores for all 48 questions. Distribution of readability indices [Flesch–Kincaid Grade Level (FKG), Gunning Fog (FOG), Automated Readability Index (ARI), Coleman–Liau Index (CLI), and composite reading grade level], length metrics (word and sentence counts), and accuracy scores for Generative Pre-trained Transformer Thinking responses under two prompt conditions (Prompt 1 vs. Prompt 2; n = 48 questions, generating 96 total evaluated responses). Grade-level scores were significantly lower with Prompt 2, indicating improved readability, whereas no statistically significant differences in word-sentence counts and accuracy scores were observed.

in demonstrating that current-generation LLMs are generally capable of producing clinically appropriate responses. At the same time, the variability observed across different LLMs in the work of Tepe and Emekli[13] underscores the importance of ongoing, comparative evaluations to understand the strengths and limitations of emerging systems as they are applied to medical communication. Nevertheless, professional human oversight remains necessary because these models can still make mistakes or produce inappropriate statements, including biases and hallucinations. Previous studies have highlighted similar concerns about using ChatGPT in radiology.[9,30,31] For example, Haver et al.[9] examined ChatGPT's ability to simplify patient-facing information on breast cancer prevention and screening. Their work showed that the model substantially reduced grade-level requirements by several levels while also shortening responses. However, they also observed that a subset of the simplified outputs contained clinically inappropriate or inaccurate statements, underscoring the need for ongoing human oversight. In our study, when source educational guidance grouped separate classification systems—such as EUSOBI referencing both U3 and BI-RADS 3 within the same clinical recommendation—the model generalized a routine surveillance approach rather than distinguishing region-specific management pathways. Although rooted in the underlying guideline text, experts recognized this oversimplification, which resulted in lower accuracy scores (mean rating of 4.0/5) and illustrates how guideline-derived question framing can propagate regional classification ambiguities into LLM evaluations.

Our study has several limitations that should be acknowledged. First, only a single LLM was evaluated, and the assessment was conducted exclusively in English. As a result, the findings cannot be generalized to other LLM versions or multilingual contexts. Second, each question–prompt pair was submitted to GPT-5 Thinking only once; given the model's non-deterministic output, we cannot rule out that some portion of the observed readability difference reflects ordinary generation variability rather than prompt manipulation alone. Repeated-sampling designs are needed to quantify this in future work. Third, although strict session isolation minimized conversational memory contamination across all steps, 33 of the 48 items were generated from SIN texts using the same LLM family that ultimately answered them. This design carries a potential risk of model-family structural affinity, which does not affect the 15 verbatim EUSOBI FAQs. Future studies should evaluate model responses against question sets derived independently by human panels or actual patients. Fourth, although all generated outputs were presented to the evaluating radiologists in an anonymized and randomized order without prompt metadata, complete evaluator blinding was inherently constrained by stylistic differences. The distinct conversational framing, structural simplification, and tone characteristic of patient-focused prompts may have allowed evaluators to infer the underlying prompt condition, potentially introducing stylistic recognition bias during qualitative scoring. Fifth, some responses included country- or context-dependent details (e.g., screening age ranges, follow-up intervals, regulatory references) that may not be accurate or applicable in all health systems; our accuracy rating did not separately flag such jurisdiction-dependent content. Sixth, readability measures reflect linguistic complexity rather than accurate comprehension. Although such indicators are well known, they do not directly assess how and to what extent patients actually understand the information, its cultural appropriateness, actionability, or real-life educational value. Therefore, they cannot replace direct evaluations of how patients understand or use the information. Seventh, although two radiologists with extensive experience independently reviewed the accuracy of each response, patient feedback was not included, which limits our ability to comment on whether patients were satisfied with the answers they received. Finally, because generation parameters and model snapshot versions cannot be controlled or retrieved through the standard web interface, exact deterministic replication of the generated text is inherently limited, as is typical for web-based LLM evaluations. Consequently, future assessments may be needed to verify whether their performance remains consistent.

Despite these limitations, our study has several important strengths. Most importantly, the evaluation was conducted within a transparent, well-documented, and auditable framework that used peer-reviewed patient information materials from EUSOBI, the leading breast imaging society in Europe. This decision ensured that the assessment was based on established clinical standards derived from EUSOBI guidelines rather than non-consensus, publicly sourced content; this benefit applies most directly to the 15 FAQ items. Furthermore, using multiple validated readability measures provided a comprehensive view of the model's linguistic complexity. Moreover, the accuracy ratings were performed independently and blindly by two radiologists, supporting the objectivity of the evaluation. Another notable strength of this study is that effect sizes across endpoints were reported with corresponding CIs rather than relying solely on *P values*, providing transparent estimation precision regarding both readability improvements and between-prompt accuracy variations. Finally, by isolating every query, both for question conversion and answer generation, in separate conversation windows, we minimized the chance that prior answer-generation interactions influenced subsequent responses. However, this does not eliminate the potential model-specific influence introduced during question generation (see Limitations).

Future research could extend upon this study in various ways. First, comparing different LLMs and multiple languages can help determine whether readability and factual accuracy remain stable despite architectural or version changes. Involving patients directly through comprehension studies is another important step in understanding how well they comprehend and trust LLM-generated explanations. Insights from such patient-centered work would reveal the actual educational impact of these tools. Finally, incorporating society-approved educational materials into clinical communication systems may help ensure that the information patients receive is both accurate and easy to understand. Together, these efforts could shape the responsible use of LLMs for patient education in breast imaging and potentially in other medical fields.

Prompt optimization substantially improved the readability of GPT-5 Thinking's breast imaging patient education content without a statistically significant reduction in expert-rated factual accuracy. Our findings underscore the model's potential as a supportive tool in patient education and health communication, although further simplification is necessary to meet widely recommended readability thresholds.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Table Link: https://d2v96fxpocvxx.cloudfront.net/68ab204c-182b-49da-b227-bc7efe058632/content-images/f36c6472-fe47-4104-97a7-debd97d186ec.pdf

References

1. Baccolini V, Isonne C, Salerno C, et al. The association between adherence to cancer screening programs and health literacy: a systematic review and meta-analysis. *Prev Med.* 2022;155:106927. [Crossref]
2. Sardanelli F, Helbich TH; European Society of Breast Imaging (EUSOBI). Mammography: EUSOBI recommendations for women's information. *Insights Imaging.* 2012;3(1):7-10. [Crossref]
3. Sardanelli F, Fallenberg EM, Clauser P, et al. Mammography: an update of the EUSOBI recommendations on information for women. *Insights Imaging.* 2017;8(1):11-18. [Crossref]
4. Mann RM, Balleyguier C, Baltzer PA, et al. Breast MRI: EUSOBI recommendations for women's information. *Eur Radiol.* 2015;25(12):3669-3678. [Crossref]
5. Evans A, Trimboli RM, Athanasiou A, et al. Breast ultrasound: recommendations for information to women and referring physicians by the European Society of Breast Imaging. *Insights Imaging.* 2018;9(4):449-461. [Crossref]
6. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230-243. [Crossref]
7. Nakaura T, Ito R, Ueda D, et al. The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. *Jpn J Radiol.* 2024;42(7):685-696. [Crossref]
8. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. *Pak J Med Sci.* 2023;39(2):605-607. [Crossref]
9. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology.* 2023;307(4):e230424. [Crossref]
10. Wong K, Fayngersh A, Traba C, Cennimo D, Kothari N, Chen S. Using ChatGPT in the development of clinical reasoning cases: a qualitative study. *Cureus.* 2024;16(5):e61438. [Crossref]
11. Soleimani M, Seyyedi N, Ayyoubzadeh SM, Kalhori SRN, Keshavarz H. Practical evaluation of ChatGPT performance for radiology report generation. *Acad Radiol.* 2024;31(12):4823-4832. [Crossref]
12. Sorin V, Barash Y, Konen E, Klang E. Deep learning for natural language processing in radiology-fundamentals and a systematic review. *J Am Coll Radiol.* 2020;17(5):639-648. [Crossref]
13. Tepe M, Emekli E. 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.* 2024;16(5):e59960. [Crossref]
14. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. *Diagn Interv Imaging.* 2023;104(6):269-274. [Crossref]
15. Rockall AG, Justich C, Helbich T, Vilgrain V. Patient communication in radiology: moving up the agenda. *Eur J Radiol.* 2022;155:110464. [Crossref]
16. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol.* 2024;30(2):80-90. [Crossref]
17. OpenAI. Introducing GPT-5. Published August 7, 2025. Accessed January 4, 2026. [Crossref]
18. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis. *Radiology.* 2024;310(3):e231593. [Crossref]
19. Pearson K, Ngo S, Ekpo E, et al. Online patient education materials related to lipoprotein(a): readability assessment. *J Med Internet Res.* 2022;24(1):e31284. [Crossref]
20. Rodriguez F, Ngo S, Baird G, Balla S, Miles R, Garg M. Readability of online patient educational materials for coronary artery calcium scans and implications for health disparities. *J Am Heart Assoc.* 2020;9(18):e017372. [Crossref]
21. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. National Technical Information Service, Springfield, Virginia;1975. [Crossref]
22. Coleman M, Liau TL. A computer readability formula designed for machine scoring. *J Appl Psychol.* 1975;60(2):283-284. [Crossref]
23. Sare A, Patel A, Kothari P, Kumar A, Patel N, Shukla PA. Readability assessment of internet-based patient education materials related to treatment options for benign prostatic hyperplasia. *Acad Radiol.* 2020;27(11):1549-1554. [Crossref]
24. Chen W, Durkin C, Huang Y, Adler B, Rust S, Lin S. Simplified readability metric drives improvement of radiology reports: an experiment on ultrasound reports at a pediatric hospital. *J Digit Imaging.* 2017;30(6):710-717. [Crossref]
25. Amin K, Doshi R, Forman HP. Large language models as a source of health information: are they patient-centered? A longitudinal analysis. *Healthc (Amst).* 2024;12(1):100731. [Crossref]
26. Gordon EB, Towbin AJ, Wingrove P, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J Am Coll Radiol.* 2024;21(2):353-359. [Crossref]
27. Haver HL, Gupta AK, Ambinder EB, et al. Evaluating the use of ChatGPT to accurately simplify patient-centered information about breast cancer prevention and screening. *Radiol Imaging Cancer.* 2024;6(2):e230086. [Crossref]
28. Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis. *Cardiovasc Intervent Radiol.* 2024;47(2):245-250. [Crossref]
29. Kaba E, Beyazal M, Çeliker FB, Yel İ, Vogl TJ. Accuracy and readability of ChatGPT on potential complications of interventional radiology procedures: AI-powered patient interviewing. *Acad Radiol.* 2025;32(3):1547-1553. [Crossref]
30. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology.* 2023;307(2):e230171. [Crossref]
31. Kao JP, Kao HT. Large language models in radiology: a technical and clinical perspective. *Eur J Radiol Artif Intell.* 2025;2:100021. [Crossref]