Utilizing Large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing

Rushabh Doshi* MSc, MPH[1]; Kanhai Amin*[,2]; Pavan Khosla BA[1]; Simar Bajaj[3], Sophie Chheang MD, MBA[4], Howard P. Forman MD, MBA[4,5,6]

*= equal contribution, co-first authors

1. Yale School of Medicine, New Haven, CT, USA
2. Yale College, New Haven, CT, USA.
3. Harvard College, Cambridge, MA, USA
4. Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA.
5. Yale School of Management, New Haven, CT, USA
6. Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA

Abstract

This paper investigates the application of Large Language Models (LLMs), specifically OpenAI's ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing, in simplifying radiology reports, thus potentially enhancing patient understanding. We examined 254 anonymized radiology reports from diverse examination types and used three different prompts to guide the LLMs' simplification processes. The resulting simplified reports were evaluated using four established readability indices. All LLMs significantly simplified the reports, but performance varied based on the prompt used and the specific model. The ChatGPT models performed best when additional context was provided (i.e., specifying user as a patient or requesting simplification at the 7th grade level). Our findings suggest that LLMs can effectively simplify radiology reports, although improvements are needed to ensure accurate clinical representation and optimal readability. These models have the potential to improve patient health literacy, patient-provider communication, and ultimately, health outcomes.

Introduction

Imaging reports are a cornerstone of medical decision-making, providing information for diagnosis, treatment planning, and monitoring disease progression. Historically, only the radiologist and referring provider accessed these reports, but the rise of telemedicine and patient portals, as well as regulatory changes, most recently the 21st Century Cures Act, have increased access to electronic health records and transformed patients' relationship with their medical information.[1–4]

Digital health literacy, defined as the degree to which a patient can obtain, process, and understand electronic information,[5] is critical for patients to fully benefit from this transformation.[6] Radiology reports, however, are filled with technical jargon, making them relatively uninterpretable to individuals without a clinical background.[7] Expanded access to these reports could thus exacerbate patient anxiety, misunderstanding, and emotional distress, particularly with abnormal findings.[8–10] Improving radiological literacy could help address these concerns, with other spillover benefits to safety and transparency,[11] shared decision-making,[12] treatment compliance,[13] and reducing health disparities.[14]

Fifteen years ago, The Joint Commission mandated that health care organizations "encourage patients' active involvement in their own care as a patient safety strategy,"[11] and a linchpin of that requirement is data transparency and accessibility. Launched in 2010, the OpenNotes program, which allowed patients to access their electronic medical records, demonstrated that 99% of patients wanted the program to continue and 85% reported that access would inform their future provider and health system choices.[15] In radiology, approaches such as leaving a summary statement at the end of the report,[16] structured templates with standardized lexicon,[17,18] and video reports[19] have all been used to improve digital health literacy. Largely underexplored are emerging artificial intelligence (AI) tools to support patient understanding.

Using deep learning techniques, large language models (LLMs), such as OpenAI's ChatGPT, Google Bard, and Microsoft Bing, have emerged as promising tools for the simplification of complex medical

information.[20,21] More specifically, these models leverage natural language processing (NLP) technologies to generate human-like text in response to a user's prompts. To date, a comparative analysis of these LLMs in radiology has not been fully explored.

In this study, we compared the performance of several popular LLMs in producing simplified reports. Our objective was to evaluate the effectiveness of LLMs and provide insights into their potential for enhancing patient health literacy and promoting better patient–provider communication.

Methods:

To investigate the efficacy of four Large Language Models (LLMs) in simplifying radiology reports, we designed a comparative study focusing on OpenAI's ChatGPT3.5 and ChatGPT4.0, Google Bard, and Microsoft Bing. Given that Bing has three conversational styles, we elected to use the precise setting over the creative or balanced settings. Our primary outcome was readability score, using an existing open-source dataset of reports.

Dataset Selection and Modification

We used the MIMIC-III database, which is a comprehensive public database from the Beth Israel Deaconess Medical Center.[22,23] A random selection of 254 anonymized reports was made to ensure representation of various examination types (MRI, CT, US (ultrasound), X-ray, Mammogram), anatomical regions, and lengths. This dataset allowed us to evaluate LLM performance across diverse clinical situations.

The reports in the datasets contained redacted information, so we altered the reports to state "Dr. Smith" where a physician name was redacted. Further, we changed redacted dates to "prior," as many reports compared findings to previous studies.

Prompt Selection

We first tested the prompt "Simplify this radiology report:" (Prompt 1). We then tested the prompt "I am a patient. Simplify this radiology report:" (Prompt 2).[24] Lastly, we tested the prompt, "Simplify this radiology report at the 7th grade level" (Prompt 3). Each prompt was followed with the radiology reports from the MIMIC-III database.

Processing Radiology Reports and Readability Assessment

Each of the 254 radiology reports were processed individually by the 4 LLMs (accessed on May 1st, 2023: ChatGPT3.5 Legacy, ChatGPT4.0, Microsoft Bing, Google Bard) generating simplified versions of the original reports for each of the three prompts. In order to standardize the outputs and ensure equal comparison, we removed all formatting, including bullet points and numbered lists, as is consistent with previous readability studies.[25,26] Ancillary information, such as "Sure I understand you would like a simplified version of your radiology report" and "please note I am not a medical professional," was also removed to focus the analysis on the clinical content.

We assessed the LLMs' ability to simplify complex radiology reports by employing four established readability indices: Gunning Fog (GF), Flesch-Kincaid Grade Level (FK), Automated Readability Index (ARI), and Coleman-Liau (CL) indices.[27] Each index outputs a score which corresponds to a reading grade level (RGL). RGL relates directly to educational attainment: an RGL of 6 corresponds to a sixth-grade level, an RGL of 12 corresponds to a high school senior level, and an RGL of 17 corresponds to a four-year college graduate level.[28–31]

As previously described,[25] we averaged the GF, FK, ARI, and CL readability scores for each output to calculate an averaged reading grade level score (aRGL). We applied the non-parametric Wilcoxon signed-rank and rank-sum tests to compare RGLs and aRGLs.

Results

We tested the LLMs with the 3 distinct prompts across 5 imaging modalities: X-ray (N=45), US (N=11), MRI (N=47), CT (N=107), and mammogram (N=33). Original radiologist reports had a median aRGL of 17.2 overall, with X-rays at 13.7, ultrasounds at 14.6, MRIs at 16.5, CTs at 18.4, and mammograms at 18.8 (Table 1). When comparing original radiologist reports, X-ray reports were significantly more readable than CT, mammogram, and MRI reports ($p<0.001$), and ultrasound reports were significantly more readable than reports for CTs and mammograms ($p<0.001$, Suppl. Fig. 2). Despite these relative differences, original X-ray and ultrasound reports were still approximately at the college RGLs.

Model	Overall Median	CT Median	X-ray Median	MRI median	US median	Mammogram median
Original Report	17.2	18.4	13.7	16.5	14.6	18.8
Prompt 1 ChatGPT3.5	10.5	10.8	10.4	10.2	10.8	9.8
Prompt 2 ChatGPT3.5	7.6	7.8	8.5	7.2	6.5	7.0
Prompt 3 ChatGPT3.5	6.7	6.9	8.0	6.6	4.6	5.5
Prompt 1 ChatGPT4.0	10.5	10.1	10.6	11.7	8.3	10.8
Prompt 2 ChatGPT4.0	8.0	7.9	8.7	8.0	8.6	7.0
Prompt 3 ChatGPT4.0	7.0	7.2	9.1	6.7	6.3	5.5
Prompt 1 Bard	9.1	8.5	12.1	8.6	7.9	9.3
Prompt 2 Bard	12.3	12.4	10.6	13.4	14.3	12.9
Prompt 3 Bard	10.1	9.9	10.3	9.6	13.8	10.6
Prompt 1 Bing	9.4	8.1	12.6	11.4	6.6	9.6
Prompt 2 Bing	12.6	12.6	12.7	12.6	11.1	12.6
Prompt 3 Bing	11.6	11.2	11.8	12.0	11.4	11.8

Table 1: *Median of the aRGL for each LLM and prompt based on examination type.*

All four LLMs significantly simplified original radiology reports from baseline complexity across all three prompts for MRI, CT, and mammogram (Figures 1-3, Suppl. Fig. 3). For X-ray and ultrasound, ChatGPT3.5, ChatGPT4.0, and Bing similarly achieved statistically significant simplification across all prompts, but Bard only simplified ultrasounds with Prompt 1 and X-rays with Prompt 2 and 3.

Prompt 1: "Simplify this radiology finding:"

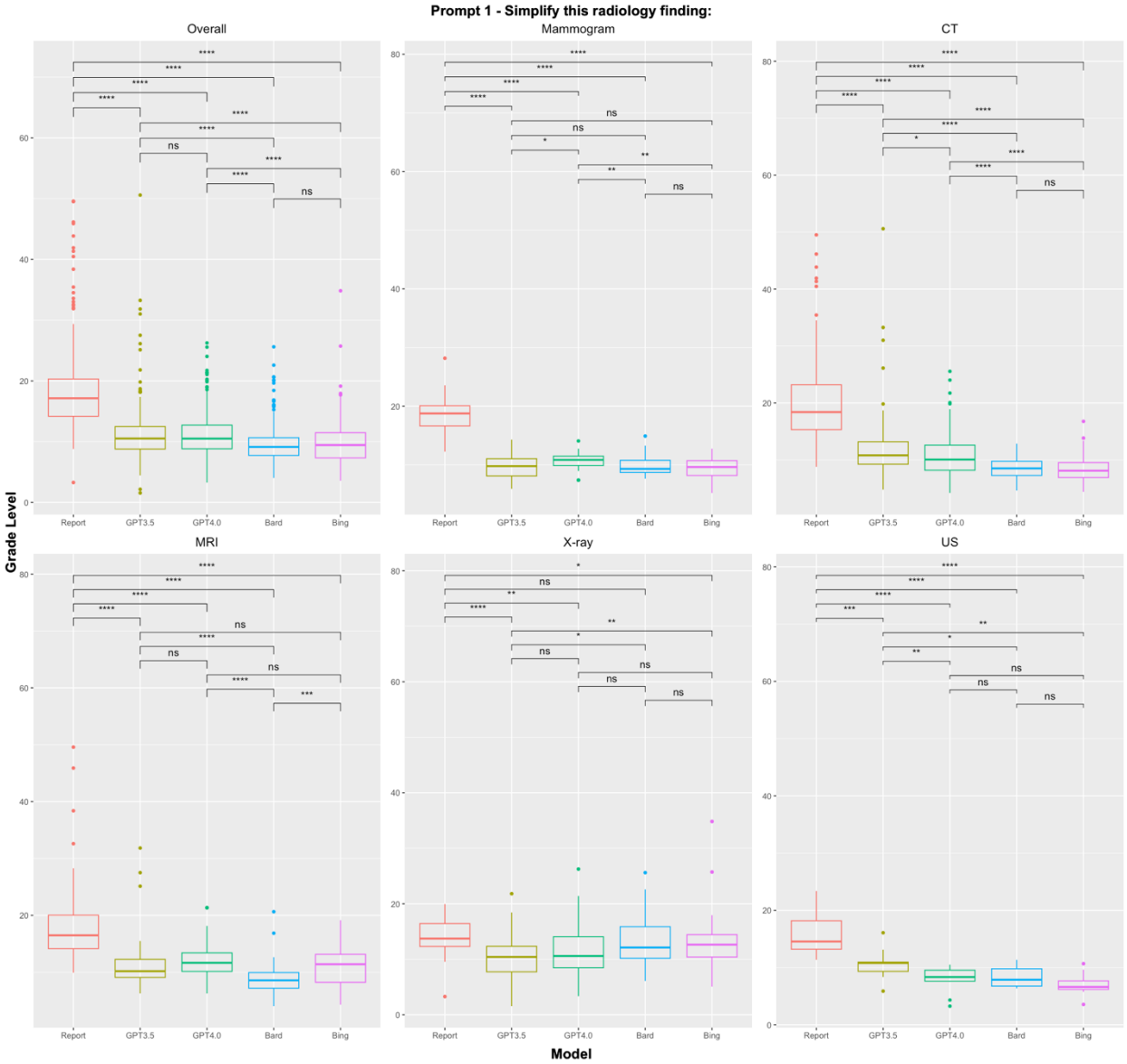


*Figure 1 Readability scores of radiologist reports and LLMs using Prompt 1 – "Simplify this radiology finding:" *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively.*

Using Prompt 1, Bing and Bard achieved significantly lower combined median aRGL (9.4 and 8.1) than ChatGPT3.5 and ChatGPT4.0 (10.5 and 10.5, p<0.0001, Figure 1). Bard and Bing otherwise performed similarly, with Bard having the lowest combined median aRGLs for MRI (8.6, p<0.001), mammogram (9.3), and overall (9.1) reports and Bing for CT (8.1) and ultrasound (6.6). With Prompt 1, ChatGPT3.5 and ChatGPT4.0 performed similarly to each another, with typically higher aRGLs than Bing and Bard. The only exception was X-rays where ChatGPT3.5 had the lowest median aRGL (10.4), significantly lower than Bard and Bing.

Prompt 2: "I am a patient. Simplify this radiology finding:"

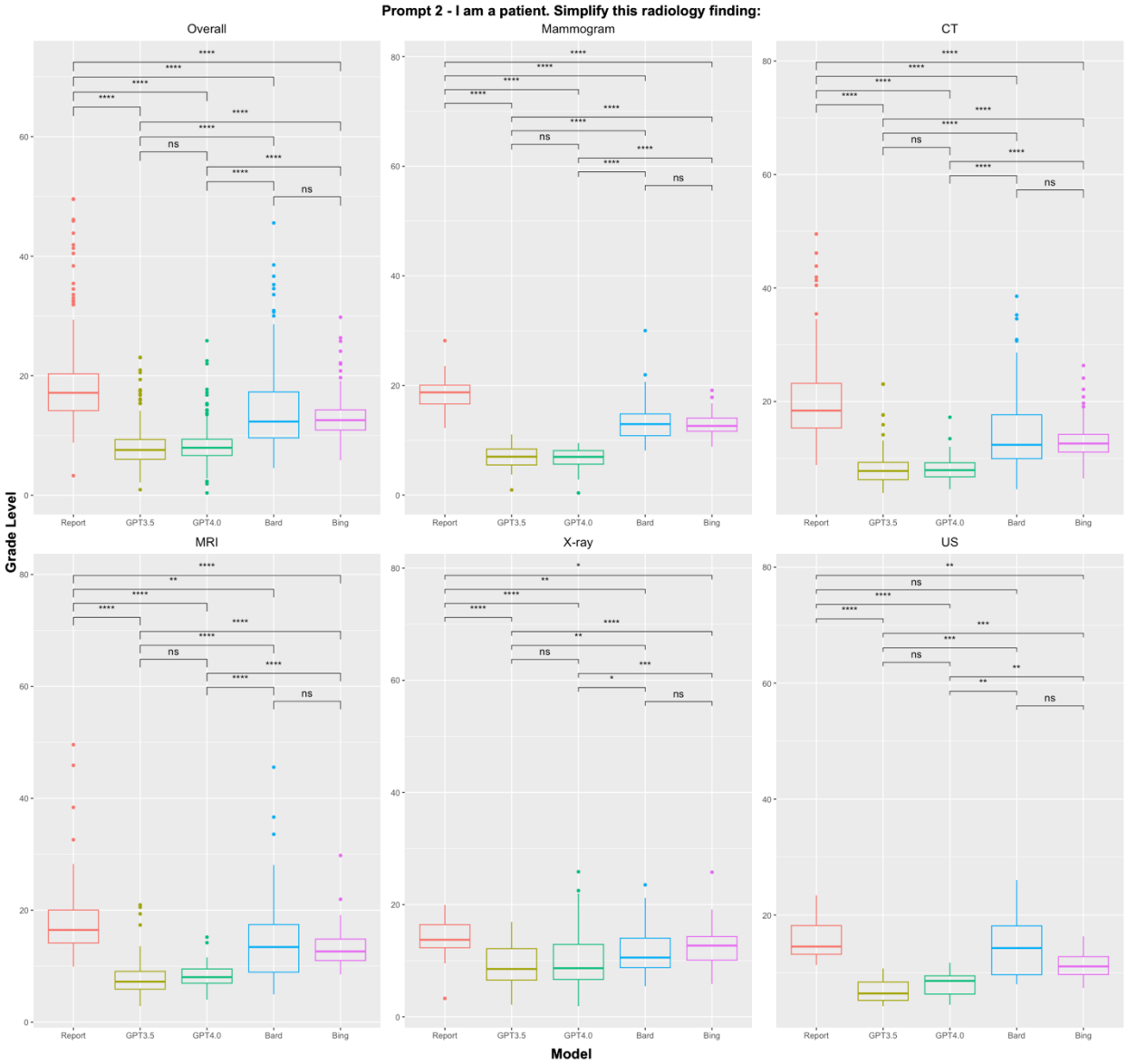


*Figure 2 Readability scores of radiologist reports and LLMs using Prompt2 – "I am a patient. Simplify this radiology finding:" *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively.*

With the added context of Prompt 2, ChatGPT3.5 and ChatGPT4.0 produced outputs with significantly lower aRGLs overall compared to Bard and Bing (p<0.0001) and for all imaging modalities tested (p<0.05, Figure 2). While there were no significant differences between ChatGPT3.5 and ChatGPT4.0, ChatGPT3.5 had the lowest median aRGL outputs for all imaging modalities (overall 7.6, CT 7.8, X-ray 8.5, MRI 7.2, US 6.5, and mammogram 7.0, Table 1).

Prompt 3: "Simplify this radiology finding at the 7th grade level:"

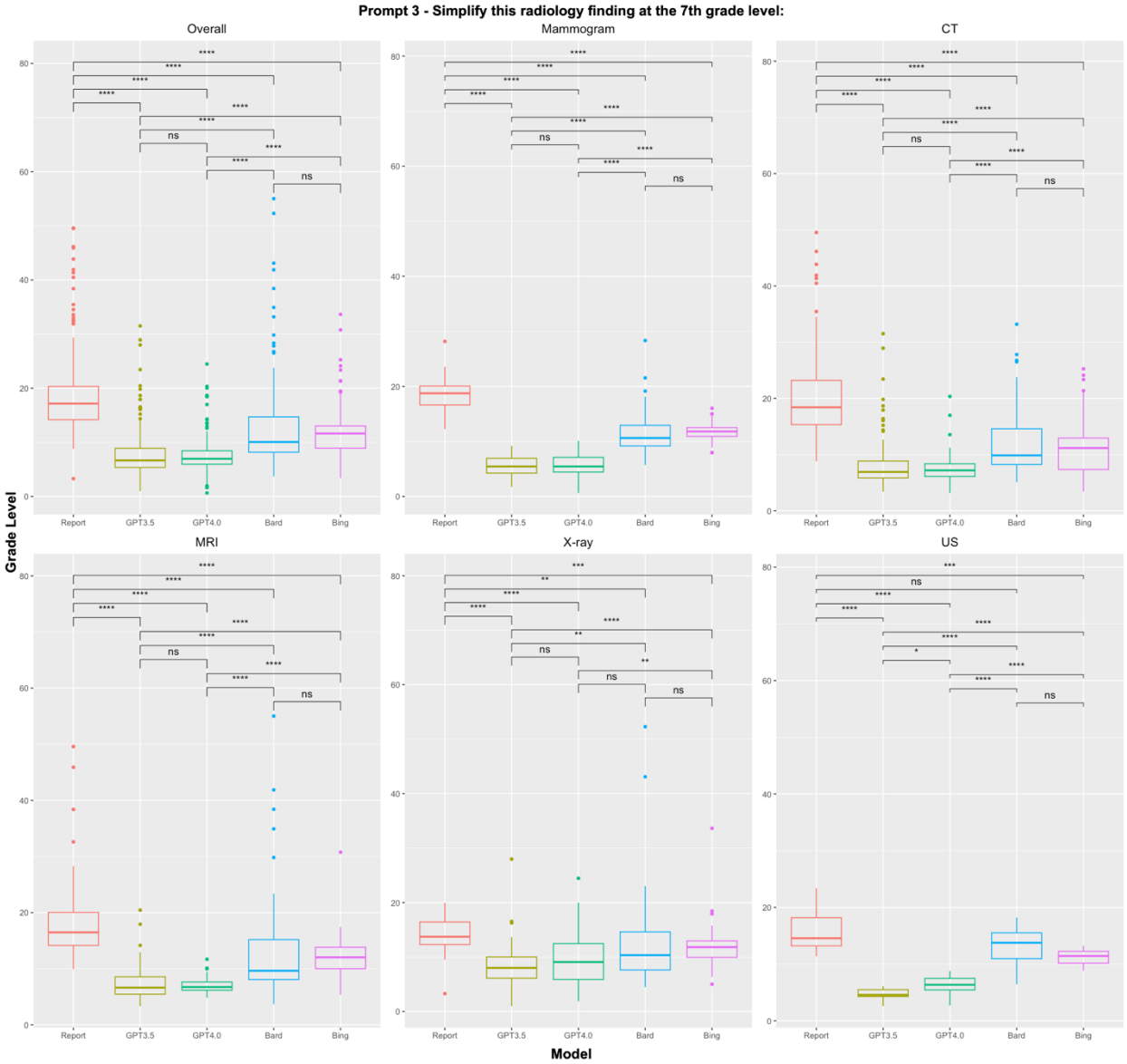


*Figure 3 Readability scores of radiologist reports and LLMs using Prompt 3 – "Simplify this radiology finding at the 7th grade level:" *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively.*

Using Prompt 3 revealed similar outcomes to Prompt 2. The ChatGPT models significantly outperformed Bard and Bing overall and across all modalities (at least p<0.01, Figure 3), except for X-rays where no difference was found between Bard and ChatGPT4. Despite the two versions performing somewhat similarly, ChatGPT3.5 again produced the lowest aRGL outputs across our analysis (overall 6.7, CT 6.9, X-ray 8.0, MRI 6.6, ultrasound 4.6, and mammogram 5.5; Table 1).

Prompt 1 vs Prompt 2 vs Prompt 3

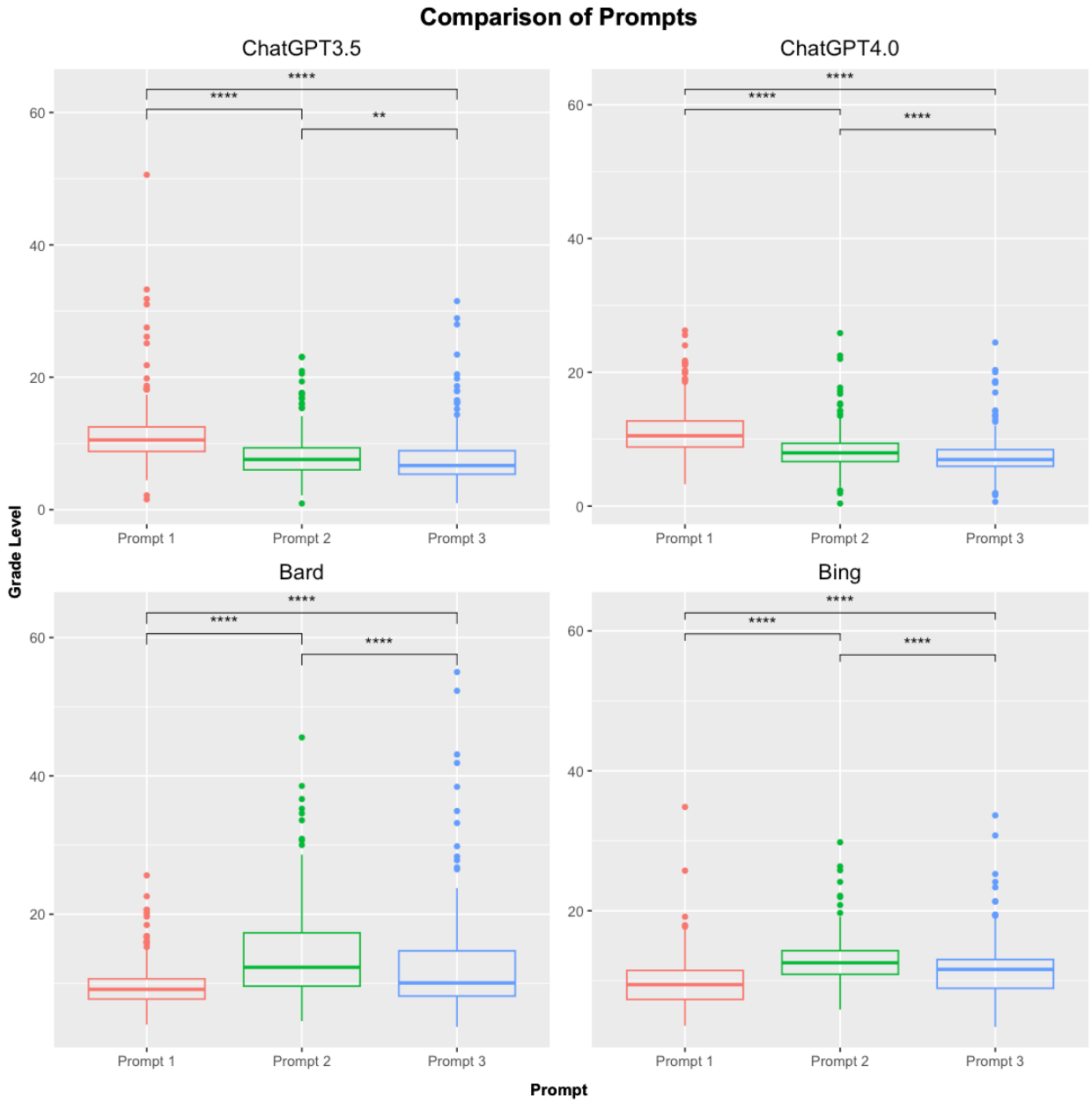


*Figure 4 Comparison of each prompt within LLM. *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively.*

Finally, we analyzed performance for each LLM across the three prompt combinations (Fig. 4). The ChatGPT models performed better at reducing aRGL with Prompt 2 and Prompt 3 than with Prompt 1 ($p<0.0001$); Prompt 3 also outperformed Prompt 2 ($p<0.01$). On the other hand, Bard and Bing performed better with Prompt 1 when compared to Prompt 2 and Prompt 3 ($p<0.0001$). We also observed that Prompt 3 outperforms Prompt 2 in producing lower aRGL outputs for Bard and Bing as well ($p<0.0001$).

Discussion

In this study, we showed that the baseline readability of radiology reports across CT, X-ray, MRI, ultrasound, and mammograms are above the college graduate level but OpenAI's ChatGPT3.5 and ChatGPT4.0, Google Bard, and Microsoft Bing can all successfully simplify these reports. The success of each of the LLMs varied, however, according to the specific prompt wording. Microsoft Bing and Google Bard performed best with a straightforward request to simplify a radiology report (Prompt 1), while the ChatGPT models performed best when provided with added context, such as the user specifying they were a patient (Prompt 2) or requesting simplification at the 7th grade level (Prompt 3).

Out of countless potential prompts that could have been tested, we focused our analysis on these three to determine how different types of context impacted readability. Prompt 1 was the simplest, specifying only that the inputted text will be a radiology report and that the LLM is tasked with simplifying it. The other two prompts offered additional context. For Prompt 3, we specified the 7th grade level because the American Medical Association and National Institutes of Health recommend that patient education materials should be written between the third- and seventh-grade levels given that the average American reads at the eighth-grade level.[18,32,33] As expected, Prompt 3 outperformed Prompt 2 across all LLMs tested, although we recognize that requesting simplification at a specific grade level is less accessible for most users than specifying that "I am a patient." Unexpectedly, however, Prompt 1 obtained the lowest aRGLs for two of the four LLMs tested, Microsoft Bing and Google Bard,—suggesting that richer context does not always equate to improved readability for every LLM.

Several explanations may underlie the observed differences in readability scores across the LLMs. For one, variations in training data and preprocessing techniques could impact the different LLMs' ability to handle the jargon, abbreviations, and numerical information found in radiology reports.[34] Furthermore, there may simply be fundamental differences in LLM architectures and algorithms that make certain models more amenable to simplifying medical information.[35] We nonetheless found the differences between Microsoft Bing and Open AI's ChatGPT models remarkable because Bing is powered by OpenAI. The finding that ChatGPT3.5 produced similar outputs to ChatGPT4.0 was also notable because it suggests that updated software does not automatically equate to improved performance, at least in regards to readability.

With patients already using these LLMs to simplify medical information,[36] providers cannot ignore how the information-sharing landscape has changed and should consider accordingly. For instance, radiologists may consider using LLMs proactively to create a patient-friendly report, inputting it into the electronic medical record alongside their original report to help alleviate patient anxiety, misunderstanding, and emotional distress.[37] Epic, Cerner, and other electronic health record companies may soon integrate LLMs into their software such that radiologists would not need to leave the interface to rely on third party tools.[38]

While LLMs demonstrate promise in helping patients better understand their radiology reports, the ultimate goal should be to strike a balance between readability and preserving clinical fidelity.[39] Indeed, excessive simplification could contribute to clinical inaccuracies and actually cause patients greater anxiety, so the role of healthcare providers in facilitating communication and understanding should not be overlooked. We believe LLMs could eventually be used as supplementary tools to aid patient-provider communication rather than a replacement for personal interaction and discussion, however, it is essential to study the accuracy and fidelity of these outputs before recommending their usage on a wider-scale.[40]

This study has limitations. For one, radiologists or medical professionals did not assess simplified outputs, so we cannot speak to the accuracy, fidelity, and clinical utility of these reports. The readability metrics used in this study are similarly limited because they are language- and structure-focused, so these measures do not necessarily capture relevance or comprehensibility from a medical perspective. Furthermore, due to the formulaic nature of these metrics, outputted RGLs were sometimes above a

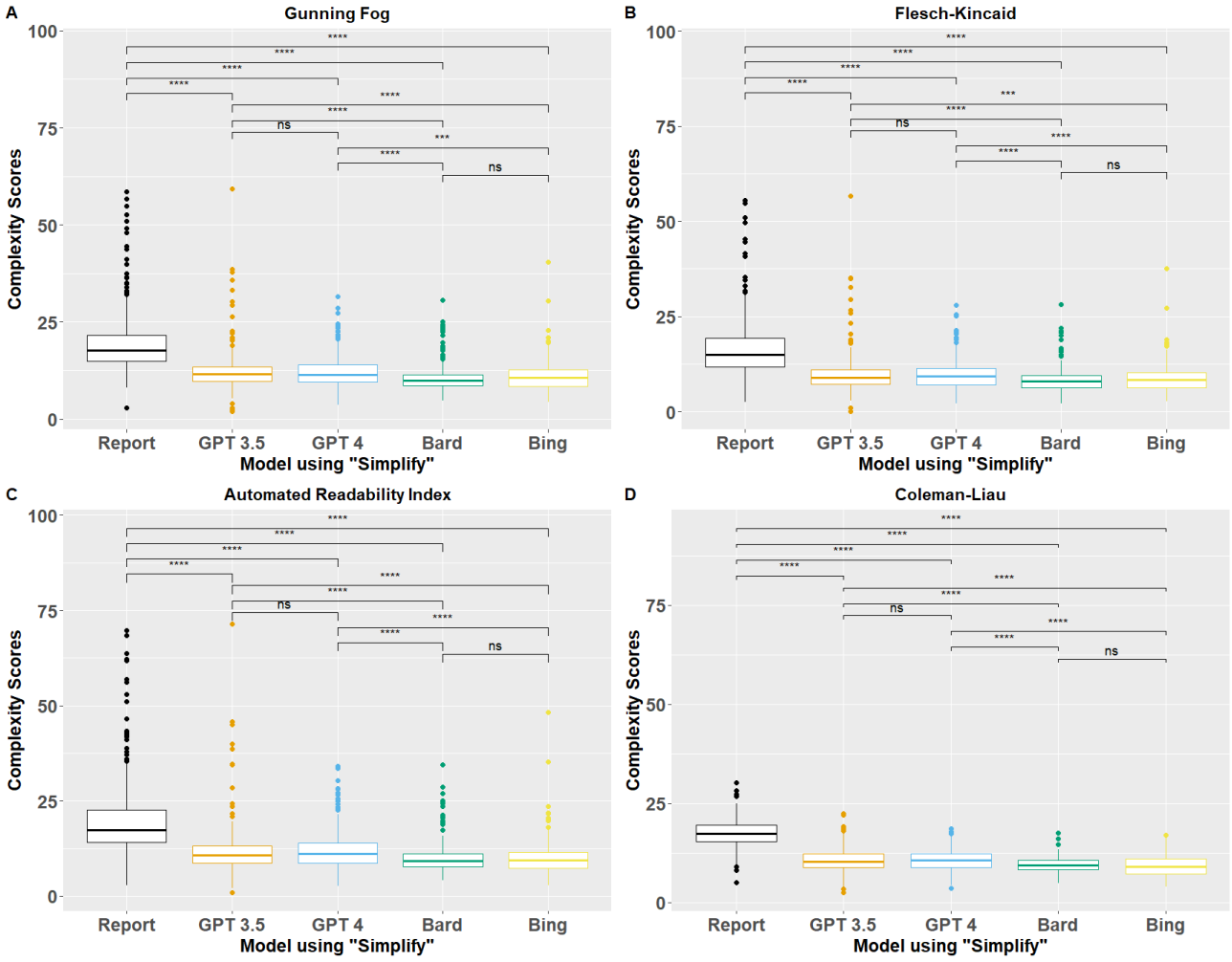
meaningful grade level (i.e., a score of 30) and thus held little interpretability on their own. In this study, we were interested in assessing the readability of reports after LLM simplification and evaluating relative differences from baseline. Finally, we extracted radiology reports from the MIMIC-III dataset, which is derived from a single hospital, and employed a cross-sectional design, which may not be ideal for capturing continuous changes in LLMs' performance. A longitudinal study design, as well as a larger, more diverse dataset, might have improved these results' validity and generalizability.

Conclusion

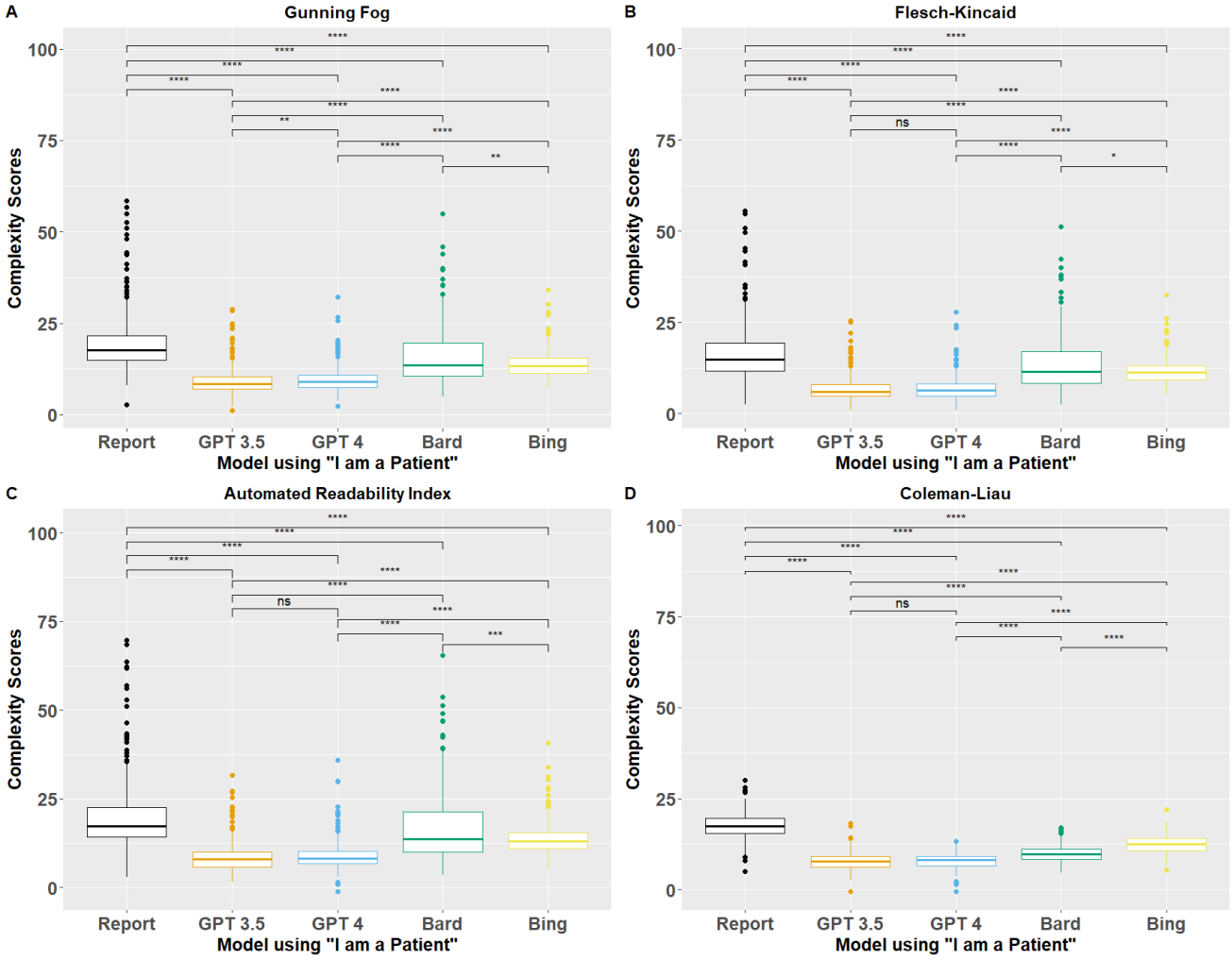
Our study highlights how radiology reports are complex medical documents that implement language and style above the college graduate reading level, but LLMs are powerful tools for simplifying these reports. Our findings should not be viewed as an endorsement for any particular LLM, instead demonstrating that each LLM tested has the ability to simplify radiology reports across modalities. Careful fine-tuning and customization for each LLM may ensure optimal simplification while maintaining the clinical integrity of the reports.

Table of Contents

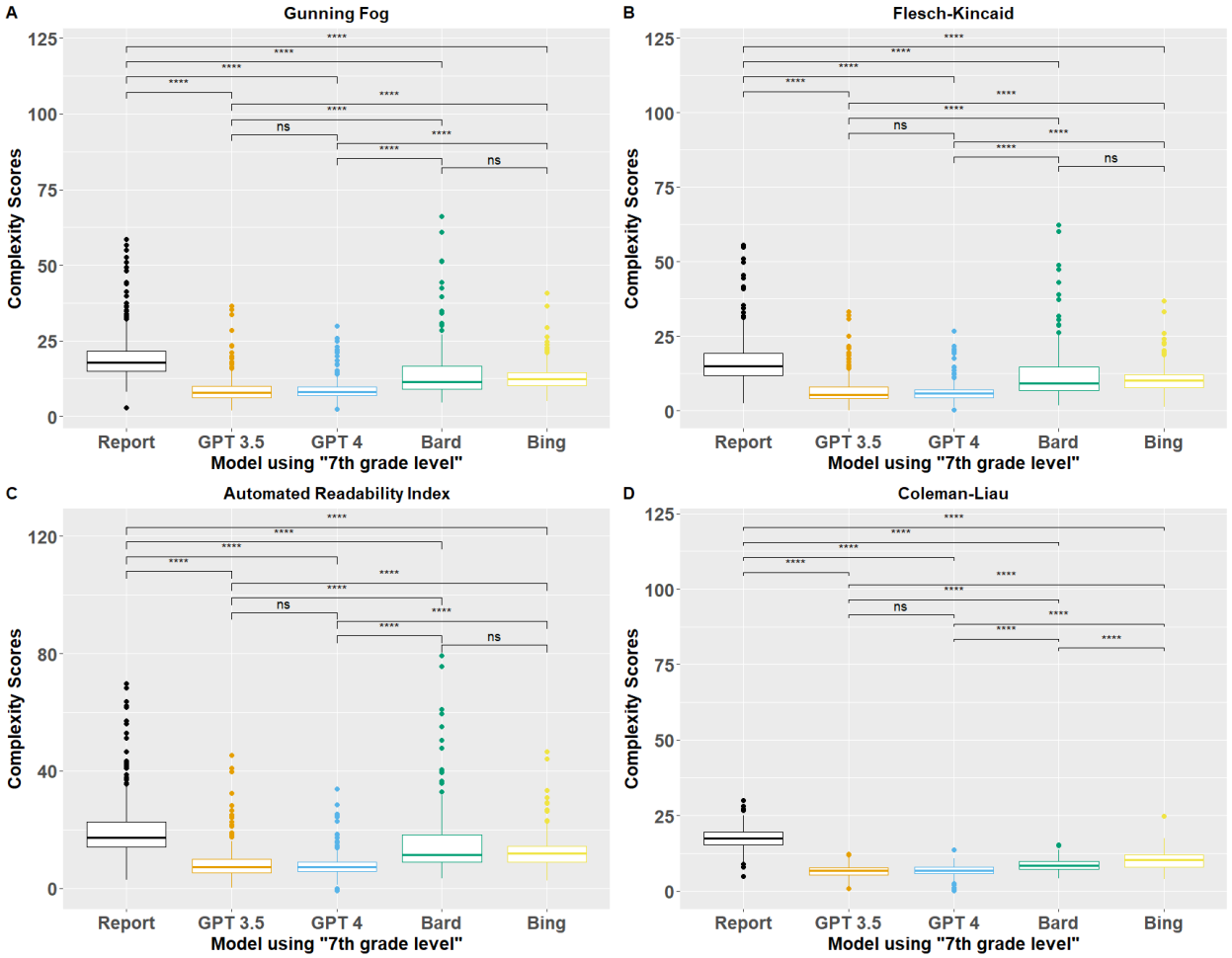
eFigure 1: GF, FK, ARI, and CL readability scores using Prompt 1. *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively



eFigure 2: GF, FK, ARI, and CL readability scores using Prompt 2. *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively



eFigure 3: GF, FK, ARI, and CL readability scores using Prompt 3. *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively



eTable 1: Median scores across LLM, prompt, modality, and readability index. *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively

Model	Total	X-Rays	MRI	Mammogram	CT	US
Original Report GF	17.7	15.12	16.52	19.94	19.48	16.4
Original Report FK	14.9	11.75	14.25	17.8	16.2	13
Original Report ARI	17.3	13	16.85	17.9	20	15.4
Original Report CL	17.35	15.77	16.83	18.16	18.18	16.73
ChatGPT-3.5 P1 GF	11.58	10.3	11.63	10.67	12.26	12
ChatGPT-3.5 P1 FK	9	8.85	8.5	8.4	9.4	8.7
ChatGPT-3.5 P1 ARI	10.8	10.6	10.55	9.8	11.4	11
ChatGPT-3.5 P1 CL	10.33	9.83	10.295	10.04	10.75	9.69
ChatGPT-3.5 P2 GF	8.48	9.175	7.83	8.04	8.4	8.28
ChatGPT-3.5 P2 FK	6	7.4	5.65	5.8	6.1	5.4
ChatGPT-3.5 P2 ARI	7.9	9.4	7.55	6.3	7.9	6.2
ChatGPT-3.5 P2 CL	7.75	7.61	7.975	7.25	7.83	5.84
ChatGPT 3.5-P3 GF	7.83	8.48	8.17	6.34	8.14	6.6
ChatGPT-3.5 P3 FK	5.4	6.65	5.4	4.2	5.8	3.6
ChatGPT-3.5 P3 ARI	7.2	8.55	7.15	5.8	7.7	3.9
ChatGPT-3.5 P3 CL	6.68	6.71	6.405	6.38	6.89	4.34
ChatGPT-4.0 P1 GF	11.35	11.435	12.35	11.34	10.71	10
ChatGPT-4.0 P1 FK	9.3	9.5	9.9	10.1	8.3	7
ChatGPT-4.0 P1 ARI	11.1	11.2	12.6	9.7	10.8	8.1
ChatGPT-4.0 P1 CL	10.62	10.36	9.93	11.72	10.41	8.28
ChatGPT-4.0 GF	9.02	9.955	9.16	8.46	9	10.96
ChatGPT-4.0 FK	6.4	7.35	6.4	5.6	6.3	8.1
ChatGPT-4.0 ARI	8.1	9.15	8.25	6.6	8.1	8.4
ChatGPT-4.0 CL	8.11	7.565	8.37	7.64	8.27	6.45
ChatGPT-4.0 GF	8.2	10.5	8.175	6.37	8.24	8.2
ChatGPT-4.0 FK	5.8	8.4	5.65	4.4	5.9	5.4
ChatGPT-4.0 ARI	7.3	10.7	7.2	5.8	7.7	6
ChatGPT-4.0 CL	6.84	6.615	6.64	6.56	6.96	5.23
Bard P1 GF	9.82	13.24	9.29	9.76	9.45	10.01
Bard P1 FK	7.9	11	7.2	8.4	7.2	6.8
Bard P1 ARI	9.3	13.85	8.55	9.4	8.9	8
Bard P1 CL	9.4	9.93	9.105	10.08	9.04	7.54
Bard P2 GF	13.6	12.25	14.745	13.73	13.8	16.72
Bard P2 FK	11.4	9.7	12.3	12.6	11.3	13.9
Bard P2 ARI	13.7	11.55	15.55	14.6	14.1	17.8
Bard P2 CL	9.8	8.995	9.685	11.04	10.04	9.93
Bard P3 GF	11.4	12.105	10.495	11.4	11.14	16.43
Bard P3 FK	9.2	9.8	8.6	9.5	9	13.8
Bard P3 ARI	11.5	11.5	10.35	11.1	11.5	16.9
Bard P3 CL	8.54	8.94	8.065	9.22	8.01	9.76
Bing P1 GF	10.64	13.485	12.545	11.17	9.26	8.9
Bing P1 FK	8.3	11	10.25	8.8	7.1	6.2
Bing P1 ARI	9.4	12.55	11.2	8.3	8.3	5.3
Bing P1 CL	9.03	11.32	10.62	9.97	7.99	5.85
Bing P2 GF	13.35	13.975	13.985	13.11	13.28	12.7
Bing P2 FK	11.2	11.8	11.35	11.9	10.8	9.8
Bing P2 ARI	13.1	13.2	13.9	12.5	13.1	10.8
Bing P2 CL	12.53	11.485	12.33	12.93	13.23	10.62
Bing P3 GF	12.44	12.17	13.1	12.86	11.75	12.39
Bing P3 FK	10.2	10.85	10.55	11.1	9.7	10.3
Bing P3 ARI	12	12.05	12.5	11.6	11.6	11.8
Bing P3 CL	10.44	9.69	11.26	11.56	9.75	11.71

eTable 2: Comparison of each modality within each prompt and LLM.

Model	Type	CT	US	Mammogram	MRI	XRAY
Report	CT	-				
	US	US (*)	-			
	Mammogram	Mammogram	US (*)	-		
	MRI	MRI (*)	US	Mammogram	-	
	XRAY	XRAY (****)	XRAY	XRAY (****)	XRAY (***)	-
P1 ChatGPT-3.5	CT	-				
	US	US	-			
	Mammogram	Mammogram (**)	Mammogram	-		
	MRI	MRI	US	Mammogram	-	
	XRAY	XRAY	US	Mammogram	XRAY	-
P2 ChatGPT-3.5	CT	-				
	US	US	-			
	Mammogram	Mammogram	US	-		
	MRI	MRI	US	Mammogram	-	
	XRAY	CT	US	Mammogram (*)	MRI	-
P3 ChatGPT-3.5	CT	-				
	US	US (****)	-			
	Mammogram	Mammogram (****)	US	-		
	MRI	MRI	US (***)	Mammogram (**)	-	
	XRAY	CT	US (***)	Mammogram (****)	MRI	-
P1 ChatGPT-4.0	CT	-				
	US	US (*)	-			
	Mammogram	Mammogram	US (***)	-		
	MRI	CT (*)	US (****)	Mammogram	-	
	XRAY	CT	US (**)	Mammogram	XRAY	-
P2 ChatGPT-4.0	CT	-				
	US	CT	-			
	Mammogram	Mammogram (**)	Mammogram	-		
	MRI	CT	US	Mammogram (**)	-	
	XRAY	CT	US	Mammogram (**)	MRI	-
P3 ChatGPT-4.0	CT	-				
	US	US	-			
	Mammogram	Mammogram (****)	Mammogram	-		
	MRI	MRI	US	Mammogram (***)	-	
	XRAY	CT (*)	US (*)	Mammogram (****)	MRI (*)	-
P1 Bard	CT	-				
	US	US	-			
	Mammogram	CT (**)	US (*)	-		
	MRI	CT	US	MRI (**)	-	
	XRAY	CT (****)	US (***)	Mammogram (****)	MRI (****)	-
P2 Bard	CT	-				
	US	CT	-			
	Mammogram	Mammogram	Mammogram	-		
	MRI	CT	US	Mammogram	-	
	XRAY	XRAY (*)	XRAY	XRAY (*)	XRAY	-
P3 Bard	CT	-				
	US	CT	-			
	Mammogram	Mammogram	Mammogram	-		
	MRI	CT	US	Mammogram	-	
	XRAY	CT	XRAY	Mammogram	XRAY	-
P1 Bing	CT	-				
	US	US (*)	-			
	Mammogram	CT (*)	US (**)	-		
	MRI	CT (****)	US (**)	Mammogram (*)	-	
	XRAY	CT (****)	US (****)	Mammogram (****)	MRI (*)	-
P2 Bing	CT	-				
	US	US	-			
	Mammogram	Mammogram	US	-		
	MRI	CT	US	Mammogram	-	
	XRAY	XRAY	US	XRAY	XRAY	-
P3 Bing	CT	-				
	US	CT	-			
	Mammogram	CT	US	-		
	MRI	CT	US	Mammogram	-	
	XRAY	CT	US	Mammogram	XRAY	-

Description: Modality listed in matrix represents modality with lower median. Significance levels are shown. *, **, ***, **** correspond to $p<0.05$, $p<0.01$, $p<0.001$, and $p<0.0001$, respectively

eTable 3: Comparison of each prompt and LLM combination within modality.

Description: P1- Prompt 1, P2 – Prompt 2, P3 – Prompt 3. Combination listed in matrix has lower median; significance levels are shown. *, **, ***, **** correspond to p<0.05, p<0.01, p<0.001, and p<0.0001, respectively

Model	Original	P1 ChatGPT-3.5	P2 ChatGPT-3.5	P3 ChatGPT-3.5	P1 ChatGPT-4.0	P2 ChatGPT-4.0	P3 ChatGPT-4.0	P1 Bard	P2 Bard	P3 Bard	P1 Bing	P2 Bing	P3 Bing
All													
P1 ChatGPT3.5	P1 3.5 (****)	-											
P2 ChatGPT3.5	P2 3.5 (****)	P2 3.5 (****)	-										
P3 ChatGPT3.5	P3 3.5 (****)	P3 3.5 (****)	P3 3.5 (**)	-									
P1 ChatGPT4.0	P1 4.0 (****)	P1 4.0	P2 3.5 (****)	P3 3.5 (****)	-								
P2 ChatGPT4.0	P2 4.0 (****)	P2 4.0 (****)	P2 3.5 (*)	P3 3.5 (****)	P2 4.0 (****)	-							
P3 ChatGPT4.0	P3 4.0 (****)	P3 4.0 (****)	P3 4.0 (**)	P3 4.0	P3 4.0 (****)	P3 4.0 (****)	-						
P1 Bard	P1 Bard (****)	P1 Bard (****)	P2 3.5 (****)	P3 3.5 (****)	P1 Bard (****)	P2 4.0 (****)	P3 4.0 (****)	-					
P2 Bard	P2 Bard (****)	P1 3.5 (****)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (****)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	-				
P3 Bard	P3 Bard (****)	P1 3.5	P2 3.5 (****)	P3 3.5 (****)	P1 4.0	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P3 Bard (****)	-			
P1 Bing	P1 Bing (****)	P1 Bing (****)	P2 3.5 (****)	P3 3.5 (****)	P1 Bing (****)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard	P1 Bing (****)	P1 Bing (****)	-		
P2 Bing	P2 Bing (****)	P1 3.5 (****)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (****)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P2 Bing	P3 Bard (****)	P1 Bing (****)	-	
P3 Bing	P3 Bing (****)	P1 3.5	P2 3.5 (****)	P3 3.5 (****)	P1 4.0	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P3 Bing (****)	P3 Bing	P1 Bing (****)	P3 Bing (****)	-
CT													
Original	-												
P1 ChatGPT3.5	P1 3.5 (****)	-											
P2 ChatGPT3.5	P2 3.5 (****)	P2 3.5 (****)	-										
P3 ChatGPT3.5	P3 3.5 (****)	P3 3.5 (****)	P2 3.5	-									
P1 ChatGPT4.0	P1 4.0 (****)	P1 4.0 (**)	P2 3.5 (****)	P3 3.5 (****)	-								
P2 ChatGPT4.0	P2 4.0 (****)	P2 4.0 (****)	P2 4.0	P2 4.0	P2 4.0 (****)	-							
P3 ChatGPT4.0	P3 4.0 (****)	P3 4.0 (****)	P3 4.0 (*)	P3 4.0	P3 4.0 (****)	P3 4.0 (***)	-						
P1 Bard	P1 Bard (****)	P1 Bard (****)	P2 3.5 (*)	P3 3.5 (**)	P1 Bard (****)	P2 4.0 (*)	P3 4.0 (****)	-					
P2 Bard	P2 Bard (****)	P1 3.5 (****)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (****)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	-				
P3 Bard	P3 Bard (****)	P3 Bard	P2 3.5 (****)	P3 3.5 (****)	P1 4.0	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P3 Bard (****)	-			
P1 Bing	P1 Bing (****)	P1 Bing (****)	P2 3.5	P3 3.5	P1 Bing (****)	P2 4.0	P3 4.0 (****)	P1 Bing	P1 Bing (****)	P1 Bing (****)	-		
P2 Bing	P2 Bing (****)	P1 3.5 (****)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (****)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P2 Bing	P3 Bard (**)	P1 Bing (****)	-	
P3 Bing	P3 Bing (****)	P3 Bing	P2 3.5 (****)	P3 3.5 (****)	P1 4.0	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P3 Bing (****)	P3 Bing	P1 Bing (****)	P3 Bing (****)	-
US													
Original	-												
P1 ChatGPT3.5	P1 3.5 (***)	-											
P2 ChatGPT3.5	P2 3.5 (***)	P2 3.5 (**)	-										
P3 ChatGPT3.5	P3 3.5 (***)	P3 3.5 (***)	P3 3.5 (**)	-									
P1 ChatGPT4.0	P1 4.0 (***)	P1 4.0 (**)	P2 3.5	P3 3.5 (**)	-								
P2 ChatGPT4.0	P2 4.0 (**)	P2 4.0 (*)	P2 3.5	P3 3.5 (**)	P1 4.0	-							
P3 ChatGPT4.0	P3 4.0 (***)	P3 4.0 (***)	P3 4.0	P3 3.5 (*)	P3 4.0 (*)	P3 4.0	-						
P1 Bard	P1 Bard (***)	P1 Bard (*)	P2 3.5 (*)	P3 3.5 (***)	P1 4.0	P2 4.0	P3 4.0 (**)	-					
P2 Bard	P2 Bard	P1 3.5 (*)	P2 3.5 (**)	P3 3.5 (***)	P1 4.0 (**)	P2 4.0 (**)	P3 4.0 (***)	P1 Bard (**)	-				
P3 Bard	P3 Bard	P1 3.5	P2 3.5 (***)	P3 3.5 (***)	P1 4.0 (**)	P2 4.0 (**)	P3 4.0 (**)	P1 Bard (**)	P3 Bard	-			
P1 Bing	P1 Bing (***)	P1 Bing (**)	P2 3.5	P3 3.5 (***)	P1 Bing	P1 Bing	P3 4.0	P1 Bing (*)	P1 Bing (***)	P1 Bing (**)	-		
P2 Bing	P2 Bing (***)	P1 3.5	P2 3.5 (**)	P3 3.5 (***)	P1 4.0 (*)	P2 4.0 (**)	P3 4.0 (***)	P1 Bard (**)	P2 Bing	P2 Bing	P1 Bing (**)	-	
P3 Bing	P3 Bing (***)	P1 3.5	P2 3.5 (***)	P3 3.5 (***)	P1 4.0 (***)	P2 4.0 (*)	P3 4.0 (***)	P1 Bard (**)	P3 Bing	P3 Bing	P1 Bing (**)	P3 Bing	-
Mammogram													
Original	-												
P1 ChatGPT3.5	P1 3.5 (****)	-											
P2 ChatGPT3.5	P2 3.5 (****)	P2 3.5 (****)	-										
P3 ChatGPT3.5	P3 3.5 (****)	P3 3.5 (****)	P3 3.5 (**)	-									
P1 ChatGPT4.0	P1 4.0 (****)	P1 3.5 (**)	P2 3.5 (****)	P3 3.5 (****)	-								
P2 ChatGPT4.0	P2 4.0 (****)	P2 4.0 (****)	P2 4.0	P3 3.5 (**)	P2 4.0 (****)	-							
P3 ChatGPT4.0	P3 4.0 (****)	P3 4.0 (****)	P3 4.0 (**)	P3 3.5	P3 4.0 (****)	P3 4.0 (**)	-						
P1 Bard	P1 Bard (****)	P1 3.5	P2 3.5 (****)	P3 3.5 (****)	P1 Bard (*)	P2 4.0 (****)	P3 4.0 (****)	-					
P2 Bard	P2 Bard (****)	P1 3.5 (****)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (***)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	-				
P3 Bard	P3 Bard (****)	P1 3.5 (*)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0	P2 4.0 (****)	P3 4.0 (****)	P1 Bard	P3 Bard (*)	-			
P1 Bing	P1 Bing (****)	P1 Bing	P2 3.5 (**)	P3 3.5 (****)	P1 Bing (*)	P2 4.0 (****)	P3 4.0 (****)	P1 Bing	P1 Bing (****)	P1 Bing (**)	-		
P2 Bing	P2 Bing (****)	P1 3.5 (****)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (****)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P2 Bing	P3 Bard	P1 Bing (****)	-	
P3 Bing	P3 Bing (****)	P1 3.5 (****)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (*)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P3 Bing (*)	P3 Bing	P1 Bing (****)	P3 Bing (*)	-
MRI													
Original	-												
P1 ChatGPT3.5	P1 3.5 (****)	-											
P2 ChatGPT3.5	P2 3.5 (****)	P2 3.5 (****)	-										
P3 ChatGPT3.5	P3 3.5 (****)	P3 3.5 (****)	P3 3.5	-									
P1 ChatGPT4.0	P1 4.0 (****)	P1 3.5 (*)	P2 3.5 (****)	P3 3.5 (****)	-								
P2 ChatGPT4.0	P2 4.0 (****)	P2 4.0 (****)	P2 3.5 (*)	P3 3.5 (*)	P2 4.0 (****)	-							
P3 ChatGPT4.0	P3 4.0 (****)	P3 4.0 (****)	P3 4.0	P3 4.0	P3 4.0 (****)	P3 4.0 (****)	-						
P1 Bard	P1 Bard (****)	P1 Bard (****)	P2 3.5 (*)	P3 3.5 (**)	P1 Bard (****)	P2 4.0	P3 4.0 (****)	-					
P2 Bard	P2 Bard (*)	P1 3.5 (**)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	-				
P3 Bard	P3 Bard (***)	P1 3.5	P2 3.5 (***)	P3 3.5 (****)	P1 4.0	P2 4.0 (***)	P3 4.0 (****)	P1 Bard (*)	P3 Bard (*)	-			
P1 Bing	P1 Bing (****)	P1 Bing	P2 3.5 (**)	P3 3.5 (****)	P1 Bing	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (**)	P1 Bing (**)	P1 Bing	-		
P2 Bing	P2 Bing (****)	P1 3.5 (***)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (*)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P2 Bing	P2 Bing	P1 Bing (**)	-	
P3 Bing	P3 Bing (****)	P1 3.5 (*)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0	P2 4.0 (****)	P3 4.0 (****)	P1 Bard (****)	P3 Bing	P3 Bing	P1 Bing	P3 Bing	-
XRAY													
Original	-												
P1 ChatGPT3.5	P1 3.5 (****)	-											
P2 ChatGPT3.5	P2 3.5 (****)	P2 3.5 (**)	-										
P3 ChatGPT3.5	P3 3.5 (****)	P3 3.5 (****)	P3 3.5	-									
P1 ChatGPT4.0	P1 4.0 (**)	P1 3.5	P2 3.5 (****)	P3 3.5 (****)	-								
P2 ChatGPT4.0	P2 4.0 (****)	P2 4.0	P2 3.5	P3 3.5 (**)	P2 4.0 (****)	-							
P3 ChatGPT4.0	P3 4.0 (****)	P3 4.0	P2 3.5	P3 3.5 (*)	P3 4.0 (***)	P3 4.0	-						
P1 Bard	P1 Bard	P1 3.5 (*)	P2 3.5 (***)	P3 3.5 (****)	P1 4.0	P2 4.0 (**)	P3 4.0 (**)	-					
P2 Bard	P2 Bard (**)	P1 3.5	P2 3.5 (**)	P3 3.5 (***)	P1 4.0	P2 4.0 (*)	P3 4.0 (*)	P2 Bard	-				
P3 Bard	P3 Bard (*)	P1 3.5	P2 3.5	P3 3.5 (**)	P1 4.0	P2 4.0	P3 4.0	P3 Bard (*)	P2 Bard	-			
P1 Bing	P1 Bing (*)	P1 3.5 (**)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (*)	P2 4.0 (****)	P3 4.0 (****)	P1 Bard	P2 Bard	P3 Bard	-		
P2 Bing	P2 Bing (*)	P1 3.5 (***)	P2 3.5 (****)	P3 3.5 (****)	P1 4.0 (*)	P2 4.0 (****)	P3 4.0 (****)	P2 Bing	P2 Bard	P3 Bard	P2 Bing	-	
P3 Bing	P3 Bing (**)	P1 3.5	P2 3.5 (***)	P3 3.5 (****)	P1 4.0	P2 4.0 (*)	P3 4.0 (*)	P3 Bing	P3 Bing	P3 Bing	P3 Bing	P3 Bing	-

REFERENCES

1. Dworkowitz A. Provider Obligations For Patient Portals Under The 21st Century Cures Act. *Health Aff Forefr*. Published online May 16, 2022. doi:10.1377/forefront.20220513.923426

2. Jain B, Bajaj SS, Stanford FC. All Infrastructure Is Health Infrastructure. *Am J Public Health*. 2022;112(1):24-26. doi:10.2105/AJPH.2021.306595

3. Lo B, Charow R, Laberge S, Bakas V, Williams L, Wiljer D. Why are Patient Portals Important in the Age of COVID-19? Reflecting on Patient and Team Experiences From a Toronto Hospital Network. *J Patient Exp*. 2022;9:23743735221112216. doi:10.1177/23743735221112216

4. Yee V, Bajaj SS, Stanford FC. Paradox of telemedicine: building or neglecting trust and equity. *Lancet Digit Health*. 2022;4(7):e480-e481. doi:10.1016/S2589-7500(22)00100-5

5. Dunn P, Hazzard E. Technology approaches to digital health literacy. *Int J Cardiol*. 2019;293:294-296. doi:10.1016/j.ijcard.2019.06.039

6. Rodriguez JA, Clark CR, Bates DW. Digital Health Equity as a Necessity in the 21st Century Cures Act Era. *JAMA*. 2020;323(23):2381-2382. doi:10.1001/jama.2020.7858

7. Bruno MA, Petscavage-Thomas JM, Mohr MJ, Bell SK, Brown SD. The “Open Letter”: Radiologists’ Reports in the Era of Patient Web Portals. *J Am Coll Radiol*. 2014;11(9):863-867. doi:10.1016/j.jacr.2014.03.014

8. Kim E, Table B, Ring D, Fatehi A, Crijns TJ. Linguistic tones in MRI reports correlate with severity of pathology for rotator cuff tendinopathy. *Arch Orthop Trauma Surg*. Published online August 23, 2022. doi:10.1007/s00402-022-04543-w

9. Bruno B, Steele S, Carbone J, Schneider K, Posk L, Rose SL. Informed or anxious: patient preferences for release of test results of increasing sensitivity on electronic patient portals. *Health Technol*. 2022;12(1):59-67. doi:10.1007/s12553-021-00628-5

10. Mehan WA, Gee MS, Egan N, Jones PE, Brink JA, Hirsch JA. Immediate Radiology Report Access: A Burden to the Ordering Provider. *Curr Probl Diagn Radiol*. 2022;51(5):712-716. doi:10.1067/j.cpradiol.2022.01.012

11. Agency for Healthcare Research and Quality. Patient Engagement and Safety. Patient Safety Network. Published September 7, 2019. Accessed May 22, 2023. https://psnet.ahrq.gov/primer/patient-engagement-and-safety

12. Waseem N, Kircher S, Feliciano JL. Information Blocking and Oncology: Implications of the 21st Century Cures Act and Open Notes. *JAMA Oncol*. 2021;7(11):1609-1610. doi:10.1001/jamaoncol.2021.3520

13. Assiri G. The Impact of Patient Access to Their Electronic Health Record on Medication Management Safety: A Narrative Review. *Saudi Pharm J SPJ*. 2022;30(3):185-194. doi:10.1016/j.jsps.2022.01.001

14. Berkman ND, Sheridan SL, Donahue KE, et al. Health literacy interventions and outcomes: an updated systematic review. *Evid ReportTechnology Assess*. 2011;(199):1-941.

15. Esch T, Mejilla R, Anselmo M, Podtschaske B, Delbanco T, Walker J. Engaging patients through open notes: an evaluation using mixed methods. *BMJ Open*. 2016;6(1):e010034. doi:10.1136/bmjopen-2015-010034

16. Kadom N, Tamasi S, Vey BL, et al. Info-RADS: Adding a Message for Patients in Radiology Reports. *J Am Coll Radiol*. 2021;18(1):128-132. doi:10.1016/j.jacr.2020.09.049

17. Panicek DM, Hricak H. How Sure Are You, Doctor? A Standardized Lexicon to Describe the Radiologist's Level of Certainty. *AJR Am J Roentgenol*. 2016;207(1):2-3. doi:10.2214/AJR.15.15895

18. Vincoff NS, Barish MA, Grimaldi G. The patient-friendly radiology report: history, evolution, challenges and opportunities. *Clin Imaging*. 2022;89:128-135. doi:10.1016/j.clinimag.2022.06.018

19. Recht MP, Westerhoff M, Doshi AM, et al. Video Radiology Reports: A Valuable Tool to Improve Patient-Centered Radiology. *Am J Roentgenol*. 2022;219(3):509-519. doi:10.2214/AJR.22.27512

20. Lyu Q, Tan J, Zapadka ME, et al. Translating Radiology Reports into Plain Language using ChatGPT and GPT-4 with Prompt Learning: Promising Results, Limitations, and Potential. Published online March 28, 2023. doi:10.48550/arXiv.2303.09038

21. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health*. 2023;5(4):e179-e181. doi:10.1016/S2589-7500(23)00048-1

22. Johnson, Alistair, Pollard, Tom, Mark, Roger. MIMIC-III Clinical Database. Published online September 4, 2016. doi:10.13026/C2XW26

23. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035. doi:10.1038/sdata.2016.35

24. Ahn S. The impending impacts of large language models on medical education. *Korean J Med Educ*. 2023;35(1):103-107. doi:10.3946/kjme.2023.253

25. 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. doi:10.2196/31284

26. 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. doi:10.1161/JAHA.120.017372

27. 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. doi:10.1007/s10278-017-9972-7

28. Habeeb A. How readable and reliable is online patient information on chronic rhinosinusitis? *J Laryngol Otol*. 2021;135(7):644-647. doi:10.1017/S0022215121001559

29. Kincaid JP, Fishburne Jr, Robert P. R, Richard L. C, Brad S. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel:* Defense Technical Information Center; 1975. doi:10.21236/ADA006655

30. Coleman M, Liau TL. A computer readability formula designed for machine scoring. *J Appl Psychol*. 1975;60:283-284. doi:10.1037/h0076540

31. 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. doi:10.1016/j.acra.2019.11.020

32. Weiss BD. *Health Literacy and Patient Safety: Help Patients Understand: Manual for Clinicians*. AMA Foundation; 2007.

33. Hansberry DR, Agarwal N, Baker SR. Health literacy and online educational resources: an opportunity to educate patients. *AJR Am J Roentgenol*. 2015;204(1):111-116. doi:10.2214/AJR.14.13086

34. Zhao WX, Zhou K, Li J, et al. A Survey of Large Language Models. Published online May 7, 2023. Accessed May 22, 2023. http://arxiv.org/abs/2303.18223

35. Fan L, Li L, Ma Z, Lee S, Yu H, Hemphill L. A Bibliometric Review of Large Language Models Research from 2017 to 2023. Published online April 3, 2023. Accessed May 22, 2023. http://arxiv.org/abs/2304.02020

36. Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT Answers Common Patient Questions About Colonoscopy. *Gastroenterology*. Published online May 5, 2023. doi:10.1053/j.gastro.2023.04.033

37. Mezrich JL, Jin G, Lye C, Yousman L, Forman HP. Patient Electronic Access to Final Radiology Reports: What Is the Current Standard of Practice, and Is an Embargo Period Appropriate? *Radiology*. 2021;300(1):187-189. doi:10.1148/radiol.2021204382

38. Landi H. HIMSS23: Epic taps Microsoft to integrate generative AI into EHRs with Stanford, UC San Diego as early adopters. Fierce Healthcare. Published April 17, 2023. Accessed May 22, 2023. https://www.fiercehealthcare.com/health-tech/himss23-epic-taps-microsoft-integrate-generative-ai-ehrs-stanford-uc-san-diego-early

39. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. Published online December 30, 2022. Accessed May 22, 2023. http://arxiv.org/abs/2212.14882

40. Doshi RH, Bajaj SS, Krumholz HM. ChatGPT: Temptations of Progress. *Am J Bioeth*. 2023;23(4):6-8. doi:10.1080/15265161.2023.2180110