European Journal of Therapeutics
pISSN: 2564-7784
eISSN: 2564-7040

Eur J Ther. 2024;30(5):714-723.
https://doi.org/10.58600/eurjther2225

Original Research

Comparative Analysis of Large Language Models in Simplifying Turkish Ultrasound Reports to Enhance Patient Understanding

Yasin Celal Güneş [1,*], Turay Cesur [2], Eren Çamur [3]

[1] Department of Radiology, Kırıkkale Yuksek Ihtisas Hospital, Kırıkkale, Türkiye
[2] Department of Radiology, Ankara Mamak State Hospital, Ankara, Türkiye
[3] Department of Radiology, Ankara 29 Mayis State Hospital, Ankara, Türkiye

Received: 2024-06-05 **Accepted:** 2024-07-31 **Published Online:** 2024-08-07

Corresponding Author
Yasin Celal Güneş, MD
Address: Kirikkale Yüksek Ihtisas Hospital, Department of Radiology, Bağlarbaşı, Ahmet Ay Caddesi, 71300 Merkez/Kırıkkale, Türkiye
E-mail: gunesyasincelal@gmail.com

ABSTRACT

Objective: To evaluate and compare the abilities of Language Models (LLMs) in simplifying Turkish ultrasound (US) findings for patients.

Methods: We assessed the simplification performance of four LLMs: ChatGPT 4, Gemini 1.5 Pro, Claude 3 Opus, and Perplexity, using fifty fictional Turkish US findings. Comparison was based on Ateşman's Readability Index and word count. Three radiologists rated medical accuracy, consistency, and comprehensibility on a Likert scale from 1 to 5. Statistical tests (Friedman, Wilcoxon, and Spearman correlation) examined differences in LLMs' performance.

Results: Gemini 1.5 Pro, ChatGPT-4, and Claude 3 Opus received high Likert scores for medical accuracy, consistency, and comprehensibility (mean: 4.7–4.8). Perplexity scored significantly lower (mean: 4.1, $p<0.001$). Gemini 1.5 Pro achieved the highest readability score (mean: 61.16), followed by ChatGPT-4 (mean: 58.94) and Claude 3 Opus (mean: 51.16). Perplexity had the lowest readability score (mean: 47.01). Gemini 1.5 Pro and ChatGPT-4 used significantly more words compared to Claude 3 Opus and Perplexity ($p<0.001$). Linear correlation analysis revealed a positive correlation between word count of fictional US findings and responses generated by Gemini 1.5 Pro (correlation coefficient = 0.38, $p<0.05$) and ChatGPT-4 (correlation coefficient = 0.43, $p<0.001$).

Conclusion: This study highlights strong potential of LLMs in simplifying Turkish US findings, improving accessibility and clarity for patients. Gemini 1.5 Pro, ChatGPT-4, and Claude 3 Opus performed well, highlighting their effectiveness in healthcare communication. Further research is required to fully understand the integration of LLMs into clinical practice and their influence on patient comprehension and decision-making.

Keywords: Large Language Models, ChatGPT, Claude 3 Opus, Ultrasound, Simplify

INTRODUCTION

Large Language Models (LLMs) are AI systems designed to comprehend and generate human language [1]. ChatGPT, developed by OpenAI and launched in November 2022, stands out as a prominent example of LLMs [2]. Alongside ChatGPT, there exist various other LLMs such as Google's Gemini, Microsoft's Copilot, Anthropic's Claude, and Perplexity. Studies across different medical specialties have evaluated the performance of LLMs [3-5].

Notably, in radiology, multiple studies have shown that large language models (LLMs) can effectively structure and simplify radiology reports, as well as educate patients about interventional radiology procedures [6-8].

Ultrasound (US) is among the most frequently utilized modalities in radiology, with reports heavily reliant on medical terminology [9]. Barrat et al.'s systematic review highlighted the negative impact of medical terminology on patient anxiety and treatment perceptions [10]. This highlights the need to provide patients with imaging reports that are easy to comprehend, especially as these reports become more readily available [11].

Main Points

- This study found that large language models (LLMs9 can effectively simplify Turkish ultrasound findings for non-medical individuals, achieving high Likert scores for accuracy and comprehensibility.
- Gemini 1.5 Pro and ChatGPT 4 emerged as top performers in terms of accuracy, comprehensibility, and readability, while Claude 3 Opus performed reasonably well but with slightly lower readability. Perplexity lagged behind in accuracy and readability.
- Transforming complex medical terminology into clear, accessible language, which can empower patients with a deeper understanding of their health status and treatment choices, highlights the potential of LLMs to facilitate patient-centered communication in healthcare, as demonstrated in this study.

Simplified reports have the potential to significantly benefit patients by enhancing their understanding of their condition and promoting adherence to treatment plans [12]. Among the various readability indices for texts, Ateşman's Readability Index specifically measures the readability of Turkish texts, assessing how easily the target audience can understand them [13].

While studies comparing the performance of LLMs in simplifying radiology reports exist for different languages, there is currently no literature available on this topic for Turkish ultrasound reports. Therefore, the aim of this study is to assess and compare the performance of different LLMs in simplifying Turkish ultrasound reports into languages understandable by patients.

MATERIALS AND METHODS

Study Design

The study conducted an assessment and comparison of several LLMs including ChatGPT 4 (https://chat.openai.com), Gemini 1.5 Pro (https://gemini.google.com), Claude 3 Opus (https://claude.ai), and Perplexity (https://perplexity.ai) to simplify Turkish US findings. Our study exclusively utilized fictional US findings and did not involve actual radiology reports, thereby exempting it from the need for ethical board approval due to the absence of real patient information. The study adhered to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines for its design and implementation [14].

Data Collection and Prompt Design

Radiologist 1 (Y.C.G.) created 50 fictional Turkish US findings used in radiology reports. Care was taken to ensure these findings were common in daily practice and portrayed realistically. The designed findings were entered into each LLM via their respective websites following the prompt, "I will write the findings from the MRI report below. Please explain them in a way that someone without a medical background can understand." in Turkish (Figure 1). Each finding was processed in a new window with default settings used for each model. The study was conducted in April 2024. All findings are shown in Table 1. The study's workflow is illustrated in Figure 2.

Table 1. Common Fictional Ultrasound Report Sentences

The liver parenchymal echo pattern appears coarse and granular.
An echo increase consistent with grade 2 steatosis is observed in the liver.
A hyperechoic lesion, primarily suggestive of a hemangioma, measuring 10 mm in segment 4 of the liver, is noted.
The gallbladder appears contracted.
Multiple stones, with the largest measuring 1 cm in diameter, are observed in the gallbladder.
The gallbladder is hydropic.
The gallbladder is hydropic, with increased wall thickness diffusely. Millimetric stones and biliary sludge are present within the gallbladder.
Biliary sludge is observed within the gallbladder.
Diffuse, millimetric hyperechoic lesions causing posterior comet-tail artifacts on the gallbladder wall are noted. The findings are suggestive of diffuse adenomyomatosis.
Phrygian cap appearance is observed in the gallbladder.
Intrahepatic bile ducts are dilated in both lobes.
Features consistent with acute cholecystitis are observed in the gallbladder.
A primarily polyp-like lesion measuring 7 mm in diameter, immobile with patient movement and non-vascular on ultrasound, is noted in the gallbladder, suggesting evaluation for acute appendicitis.
Pancreas and midline structures cannot be evaluated due to gas.
A cystic lesion measuring 7 mm in diameter, without solid components or septa, is observed in the pancreatic head.
The pancreatic parenchyma appears diffusely edematous, with fluid collections in the peripancreatic area.
The spleen measures 14 cm in size and appears enlarged.
An accessory spleen measuring 7 mm in diameter is observed at the splenic hilum.
A nodular lesion measuring 8 mm in diameter, demonstrating arterial vascularity on ultrasound, is observed at the splenic hilum, suggesting a splenic artery aneurysm.
Collateral vascular structures are observed at the splenic hilum.
Grade 2 increase in echogenicity is noted in the parenchyma of both kidneys.
Both kidneys appear decreased in size and parenchymal thickness.
Multiple simple cortical cysts measuring less than 1 cm in diameter are observed in both kidneys.
Fusion of the lower poles of both kidneys to the midline of the abdomen anterior to the abdominal aorta is observed, suggesting a horseshoe kidney.
A hyperechoic nodular lesion, primarily suggestive of an angiomyolipoma, measuring 8 mm in the mid-pole of the right kidney, is observed.
An isoechoic solid mass lesion with internal vascularity measuring 27x24 mm, extending exophytically in the lower segment of the right kidney, is observed. Dynamic upper abdominal CT/MRI is recommended for lesion characterization.
Grade 2 dilation of the pelvicalyceal system is noted in both kidneys.
A hyperechoic solid lesion measuring 17x12 mm, demonstrating vascularity on ultrasound, with papillary extensions into the lumen, is observed in the posterolateral wall of the bladder. Histopathological correlation is recommended.
There is diffuse thickening of the bladder wall.
There is increased trabeculation of the bladder wall.
Widespread millimetric echogenicities are observed within the bladder lumen. Clinical and laboratory evaluation is recommended for cystitis.
A submucosal edema measuring 9 mm in diameter, originating from the terminal ileum in the right lower quadrant, with a blind-ending tubular segment unresponsive to compression, is observed. Clinical and laboratory evaluation is recommended for acute appendicitis.
Omental fat tissue and bowel loops demonstrating herniation from a fascial defect measuring approximately 13 mm in the right inguinal region with valsalva maneuver are observed.
Omental fat tissue and bowel loops demonstrating herniation from a fascial defect measuring approximately 13 mm in the right inguinal region with valsalva maneuver are observed. There is no return of herniated material following the valsalva maneuver. Thickening of up to 7 mm is measured at the thickest point of the herniated bowel loops, with loss of vascularity on ultrasound examination. Fluid measuring 5 mm in depth is observed within the hernia sac.

Multiple simple cysts measuring less than 5 mm in diameter are observed in both breasts.
A solid lesion measuring 13x14 mm, parallel to the skin with regular contours and no vascularity on ultrasound examination, is observed approximately 24 mm from the nipple at 3 o'clock in the right breast (BI-RADS 3).
A hypoechoic solid lesion demonstrating spiculated margins and vascularity on ultrasound examination, measuring 13x14 mm, is observed approximately 33 mm from the nipple at 6 o'clock in the right breast (BI-RADS 5).
An echogenic hilum lymph node measuring 13 mm in diameter is observed in the right axilla.
An LAP with asymmetric cortical thickening measuring 17 mm in diameter, with a cortex thickness of 3.4 mm at the thickest point, is observed in the right axilla.
The thyroid parenchyma appears heterogeneous with scattered hypoechoic areas and echogenic fibrous septa.
An isoechoic nodule with microcystic degeneration areas measuring 16 mm in diameter is observed in the mid-section of the right thyroid lobe.
Multiple isoechoic nodules, with the largest measuring 8 mm in the mid-section of the right lobe and 11 mm in the lower section of the left lobe, are observed in both thyroid lobes.
A hypoechoic solid nodule with irregular margins measuring 13x13 mm in the lower section of the right thyroid lobe is observed, with microcalcifications within.
Multiple lymph nodes with echogenic hilum, measuring less than 1 cm in short diameter and with thick cortex, are observed in both cervical chains.
A cystic appearance primarily suggestive of a dominant follicle, measuring 23 mm, is observed in the right ovary.
A heterogeneous lesion with thick septa and a solid component demonstrating vascularity on ultrasound examination, measuring 23x24 mm, is observed in the left ovary. Contrast-enhanced pelvic MRI is recommended for lesion characterization.
An intramural myoma measuring 13 mm in diameter is observed in the right half of the uterine corpus.
Endometrial thickness, measured as double-wall thickness, is 17 mm and significantly increased.
A tubular structure measuring 14 mm in diameter in the right adnexal region, suggestive of a tubal structure consistent with hydrosalpinx, is observed.
A lesion measuring 17 mm in diameter with multiple septa within, primarily suggestive of a hemorrhagic cyst, is observed in the right ovary.

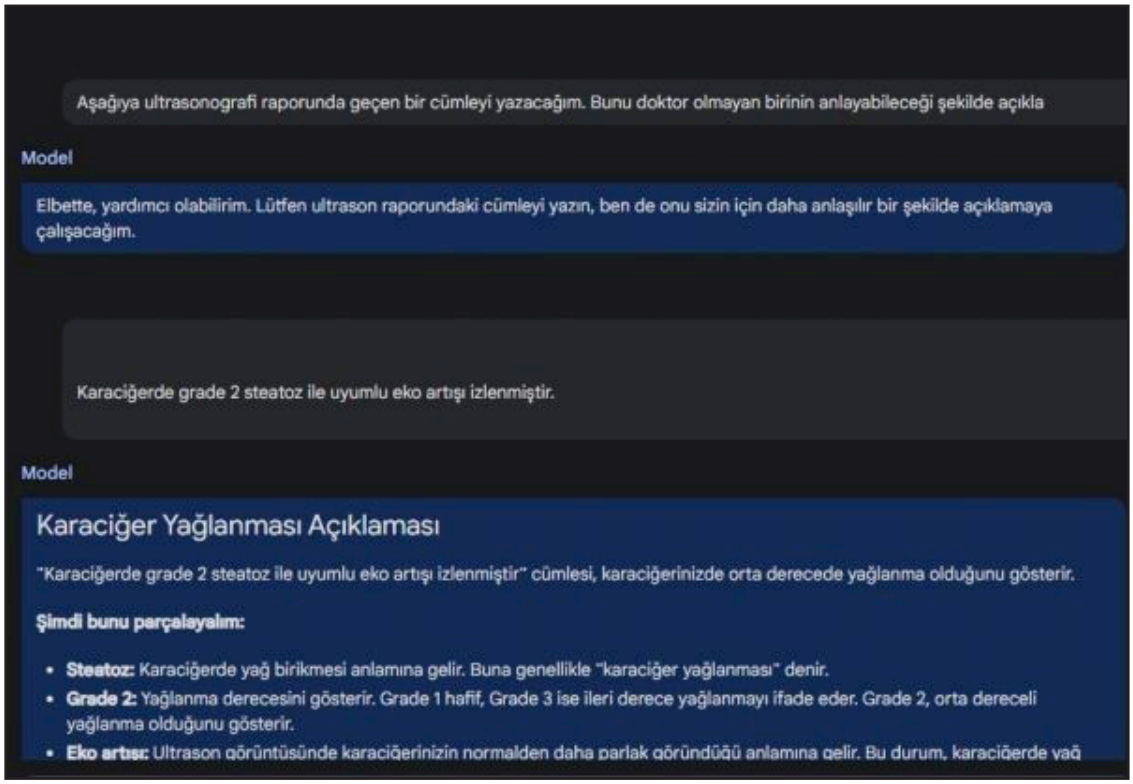

Figure 1. Demonstrates the Turkish input and output process for Turkish ultrasound findings in the study through Gemini 1.5 Pro.

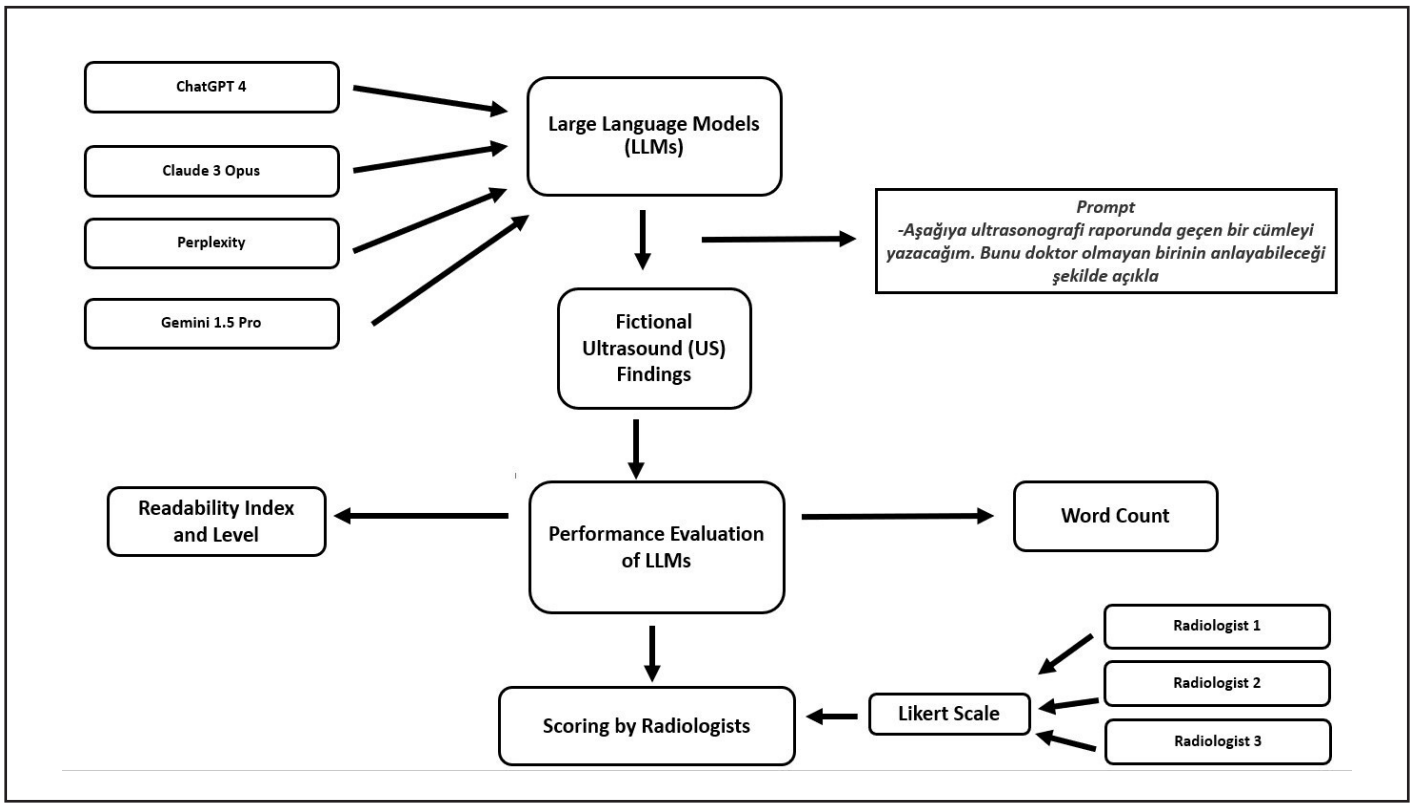

Figure 2. The study's workflow is illustrated in Figure 2.

Performance Evaluation of LLMs

The responses from the LLMs were analyzed using the Ateşman's Readability Index [198,825 - (40,175 x number of syllables/number of words) - (2,610 x number of words/number of sentences)] to determine readability levels in Turkish which has been introduced in 1997 by Ender Ateşman (Table 2) [13]. The word length used in the formula is calculated in syllables and sentence length is calculated in words. The publicly available and free website "http://okunabilirlikindeksi.com" was used for this analysis. The responses were rated jointly by three radiologist [Radiologist 1 (Y.C.G.), Radiologist 2 (T.C.), Radiologist 3 (E.Ç.), each with 6 years of experience in general radiology and certified by the European Diploma in Radiology (EDiR)] using a Likert Scale from 1 to 5 based on medical accuracy, consistency of recommendations, and comprehensibility. The number of words in each response was recorded. Additionally, the word count and Ateşman's Readability Index of the fictional US findings were compared to the responses generated by each LLM.

Statistical Analysis

Data distribution was examined using the Kolmogorov-Smirnov and Shapiro-Wilk tests, while the Levene test was utilized to assess data variance. Descriptive statistics, comprising measures such as minimum, maximum, average, median, standard deviation, interquartile range, and percentages, were then calculated. Subsequently, to identify significant relationships among quantitative data within dependent groups, both the Friedman and Wilcoxon tests were employed. Additionally, Spearman correlation analysis was conducted to investigate the linearity of correlations between quantitative data. Statistical analyses were performed using IBM SPSS Version 26.

RESULTS

Likert Scale

No statistically significant difference was found between the scores of ChatGPT 4 (mean: 4.82; median: 5.0), Gemini 1.5 Pro (mean: 4.78; median: 5.0), and Claude 3 Opus (mean: 4.68; median: 5.0) based on the Likert scale ($p>0.05$). However, the scores of ChatGPT 4, Gemini 1.5 Pro, and Claude 3 Opus were significantly higher than those of Perplexity (mean: 4.08; median: 4.0) ($p<0.001$). No difference was found between the scores of Claude 3 Opus and Gemini 1.5 Pro ($p=0.39$) (Figure 3) (Table 3).

Table 2. The Ateşman Readability Index and Its Corresponding Readability Level

Index	Readability Level
90 - 100	Easily understood by 4th grade and below students
80 - 89	Easily understood by 5th or 6th graders
70 - 79	Easily understood by 7th or 8th graders
60 - 69	Easily understood by 9th or 10th graders
50 - 59	Easily understood by 11th or 12th graders
40 - 49	Easily understood by 13th or 15th-year (associate degree) students
30 - 39	Easily understood by bachelor's degree
< 30	Easily understood by postgraduates

Table 3. Descriptive Findings of the Responses of the Large Language Models.

	Gemini 1.5 Pro	Claude 3 Opus	ChatGPT 4	Perplexity
Likert Scores*				
Minimum-Maximum	2.0-5.0	3.0-5.0	3.0-5.0	1.0-5.0
Mean ± SD	4.8 ± 0.5	4.7 ± 0.6	4.8 ± 0.5	4.1 ± 0.6
Median (IQR)	5.0 (0)	5.0 (0)	5.0 (0)	4.0 (1.0)
Ateşman Readability Index				
Minimum-Maximum	45.0 - 72.0	28.0 - 66.0	29.0-73.0	26.0 – 71.0
Mean ± SD	61.16 ± 6.66	51.16 ± 7.23	58.94 ± 9.61	46.25 ± 11.98
Median (IQR)	54.1 (14.2)	47.2 (10.3)	37.8 (7.6)	37.3 (6.2)
Readability Level				
Minimum	7.-8. Grade	9.-10. Grade	7.-8. Grade	7.-8. Grade
Maximum	Associate's Degree	Postgraduate	Postgraduate	Postgraduate
Median	11.-12. Grade	11.-12. Grade	11.-12. Grade	13.-15. Grade
Word Count				
Minimum-Maximum	77-283	69-197	57-292	40-187
Mean ± SD	166.9 ± 46.9	99.24 ± 36.3	154.9 ± 39.8	88.4 ± 26.5
Median (IQR)	171.5 (24.5)	79.0 (47.0)	157.0 (72.0)	91.5 (35.0)

*Likert Scores: In our study, the accuracy of the explanations, consistency and comprehensibility of the suggestions made by the big language models were rated on a scale of 1 to 5.

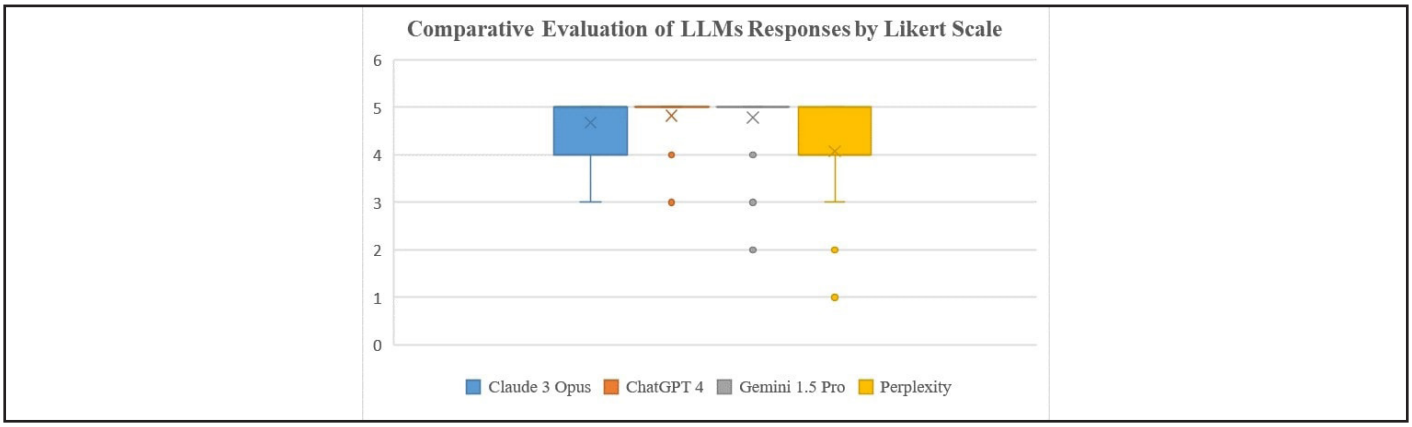

Figure 3. A box-plot displays the radiologists' consensus scores for the Large Language Models' answers, with the median score denoted by 'x' and outlying scores marked by dots.

SD: Standard Deviation, IQR: Interquartile range.

Ateşman's Readability Index

A significant difference was found between the mean scores of the LLMs based on the Ateşman's Readability Index ($p<0.001$). The highest mean score was 61.16 for Gemini 1.5 Pro, and the second-highest score was 58.94 for ChatGPT 4, which did not show a significant difference ($p=0.208$). Additionally, both Gemini 1.5 Pro and ChatGPT 4 had significantly higher Ateşman's Readability Index scores than Claude 3 Opus (51.16) and Perplexity (47.01) ($p<0.001$). Claude 3 Opus's Ateşman's Readability Index score was also significantly higher than Perplexity's ($p=0.006$) (Table 3).

Word Counts

No significant difference was found between the mean word count of Gemini 1.5 Pro (166.2) and ChatGPT 4 (154.38) ($p=0.213$). However, both Gemini 1.5 Pro and ChatGPT 4 used significantly more words than Claude 3 Opus (99.24) and Perplexity (93.48) ($p<0.001$). No significant difference was found between the word count of Claude 3 Opus and Perplexity ($p>0.05$) (Table 3).

Correlation Analysis

A linear correlation was observed between the word count of fictional US findings and the word count of responses generated by Gemini 1.5 pro (correlation coefficient = 0.38, $p<0.05$) and ChatGPT 4 (correlation coefficient = 0.43, $p<0.001$). Conversely, no correlation was found between the word count of Claude 3 Opus ($p>0.05$) and Perplexity ($p\geq0.05$).

A significant correlation was identified between the Ateşman's Readability Index of fictional US findings and the Ateşman's Readability Index of responses from Claude 3 Claude 3 Opus (correlation coefficient = 0.42, $p<0.001$), ChatGPT 4 (correlation coefficient = 0.51, $p<0.001$), and Gemini 1.5 pro (correlation coefficient = 0.45, $p<0.001$). However, no correlation was detected between the Ateşman's Readability Index of fictional US findings and Perplexity ($p>0.05$).

DISCUSSION

Our study found that LLMs effectively simplified commonly used Turkish US findings for non-medical individuals. Gemini 1.5 Pro, Claude 3 Opus, and ChatGPT 4 received high Likert scores (4.68 to 4.82 out of 5), indicating their effectiveness in conveying US findings clearly. However, Perplexity scored lower (4.1), suggesting less accuracy and comprehensibility. Gemini 1.5 Pro had the highest readability score (61.16), while ChatGPT 4 (58.94) and Claude 3 Opus (51.16) followed closely. Perplexity had the lowest average score (46.25). These results suggest that Gemini 1.5 Pro produced the easiest to understand responses, while Perplexity generated the least readable responses.

In terms of readability, all LLMs predominantly produced responses suitable for readers with at least an associate degree or higher educational background. However, there were variations in the readability levels produced by each model. Gemini 1.5 Pro and ChatGPT 4 mainly generated responses at a 7th to 8th-grade reading level, while Claude 3 Opus and Perplexity tended to produce responses at a slightly higher reading level, ranging from 9th to 12th grade.

Gemini 1.5 Pro and ChatGPT 4 provided more detailed explanations with higher word counts, while Claude 3 Opus and Perplexity had lower word counts. Overall, Gemini 1.5 Pro and ChatGPT 4 emerged as top performers in accuracy, comprehensibility, and readability, while Claude 3 Opus performed reasonably well but with slightly lower readability. Perplexity lagged behind in accuracy and readability.

The variations in performance among LLMs could be attributed to their differing design architectures. For example, Perplexity's lower accuracy and readability scores might result from its unique feature of having internet access, potentially incorporating less reliable online sources [15].

Previous studies have also demonstrated the effectiveness of LLMs in simplifying radiology reports [16-21]. Doshi et al. conducted a study showcasing the efficacy of ChatGPT 4, ChatGPT 3.5, Google Bard, and Microsoft Copilot in simplifying 750 radiology reports. These reports, spanning various modalities like ultrasound (US), CT, MRI, mammography, and X-ray, were subjected to three distinct prompts [16]. The first prompt sought a general simplification of the report, while the second prompt required the report to be simplified from a patient's perspective. The final prompt mandated the report to be simplified to a 7th-grade reading level [16]. In our study, a prompt similar to the second prompt used by Doshi was successfully employed to simplify ultrasound reports.

Haver et al. [17] demonstrated that ChatGPT could simplify responses to 25 breast cancer questions to a sixth-grade reading level. They evaluated the original and simplified responses

using the Flesch Reading Ease Index and five readability scales. Ninety-two percent of the simplified responses were deemed clinically appropriate, showing significant improvements in reading ease, readability, and word count reduction.

Chung et al. [18] utilized ChatGPT to summarize MRI reports of prostate cancer patients, customizing them for patient comprehension levels. Their study yielded fifteen summarized reports from five full MRI reports, revealing a noteworthy decrease in the median Flesch-Kincaid Grade Level (FKGL) score from 9.6 to 5.0.

Li et al. [19] showcased ChatGPT's efficacy in simplifying radiology reports to below an 8th-grade reading level across various modalities, including radiographs, ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI), demonstrating its versatility in different imaging contexts. Their study revealed that the mean report length was 164 words, accompanied by a Flesch reading ease score of 38.0 and a FKGL of 10.4.

In Lyu et al.'s study [20], radiologists evaluated ChatGPT's output for chest CT and brain MRI scans, acknowledging some missing and incorrect information. Despite these discrepancies, the overall quality score was 4.268 out of 5 on a five-point Likert scale, with minor discrepancies of 0.08 and 0.07 for missing and misinformation, respectively.

Tepe et al. [21] compared the effectiveness of ChatGPT 4, Google Bard, and Microsoft Copilot in translating CT and MRI reports into patient-friendly language, with all models achieved understandability scores above 70%. but Bard showing superior readability scores.

Our findings align with these studies, showing that ChatGPT 4, Gemini 1.5 Pro, and Claude 3 Opus effectively simplified Turkish US findings, achieving high Likert scores for accuracy and comprehensibility. However, Perplexity scored lower in these areas. Gemini 1.5 Pro and ChatGPT 4 also produced the highest readability scores, indicating their superior ability to simplify medical language.

In contrast to previous studies, our research focused specifically on Turkish US reports, providing new insights into the performance of LLMs in this context rather than English radiological reports. Therefore, we used the Ateşman's Readability Index for the Turkish language, rather than the Flesch Reading Ease Index and other readability scales.

Our study also indicates the potential of LLMs to facilitate patient-centred communication in healthcare. By transforming complex medical terminology into clear, accessible language, these systems have the potential to empower patients with a deeper understanding of their health status and treatment choices. This enhanced comprehension can facilitate more confident decision-making, heightened patient involvement, and ultimately, improved health outcomes.

Limitations

Although our study represents the first investigation into the simplification of Turkish US findings by LLMs for individuals without medical backgrounds, it has several limitations. Firstly, although the US sentences utilized in our study are commonly encountered in daily medical practice, they are synthetic in nature. To provide a more accurate assessment of LLMs' performance, real-world reports from hospital settings should be incorporated into future studies. Secondly, the sample size of sentences used in our study was limited, and they primarily focused on straightforward cases. To enhance the applicability of our findings, future research should expand the scope to include a broader range of complex sentences tailored to different anatomical regions. Thirdly, our study employed a single prompt, which may not fully capture the range of scenarios encountered in clinical practice. Further investigation into the impact of different prompts on LLM performance is warranted to better understand how varying contexts influence their effectiveness. Lastly, our study lacked data on real-life patients' satisfaction and comprehension of simplified reports. Prospective multicenter studies utilizing simplified reports generated by LLMs and assessing patients' understanding and satisfaction levels are essential for validating the practical utility of these systems in healthcare communication.

CONCLUSIONS

In conclusion, our study demonstrates that LLMs successfully simplify Turkish US findings, emphasizing their role in improving accessibility and understandability of radiological information for patients. Further research and implementation efforts are needed to fully harness the potential of LLMs in facilitating effective communication between healthcare

providers and patients.

Acknowledgments: The authors used ChatGPT, a language model based on the GPT-3.5 architecture (May 2024 Version; OpenAI; https://chat.openai.com/) to revise the grammar and English translation. The content of the publication is entirely the authors' responsibility, and the authors examined and edited it as necessary.

Informed Consent: No informed conset was required for this study.

Conflict of interest: : The authors declare that this study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding: No funding was received for this study.

Ethical Approval: Our study only included fictional ultrasound findings and did not use actual radiology reports or patient information, so it did not require ethical board approval.

Author Contributions: Conception: Y, C, G; E, Ç; T, C - Design: Y, C, G; E, Ç; T, C - Supervision: Y, C, G - Fundings: - -Materials: Y, C, G; E, Ç; T, C - Data Collection and/or Processing: Y, C, G; E, Ç; T, C - Analysis and/or Interpretation: Y, C, G; E, Ç; T, C - Literature: Y, C, G; E, Ç; T, C - Review: Y, C, G; E, Ç; T, C - Writing: Y, C, G - Critical Review: Y, C, G; E, Ç; T, C

REFERENCES

[1] Aydin Ö, Karaarslan E (2023) Is ChatGPT Leading Generative AI? What is Beyond Expectations? Academic Platform Journal of Engineering and Smart Systems 11:118-134. https://doi.org/10.21541/apjess.1293702

[2] Lee H (2023) The rise of ChatGPT: Exploring its potential in medical education. Anatomical sciences education. https://doi.org/10.1002/ase.2270

[3] Kuang Y-R, Zou M-X, Niu H-Q, Zheng B-Y, Zhang T-L, Zheng B-W (2023) ChatGPT encounters multiple opportunities and challenges in neurosurgery. International Journal of Surgery 109:2886-2891. https://doi.org/10.1097/JS9.0000000000000571

[4] Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J (2023) Challenging ChatGPT 3.5 in senology—an assessment of concordance with breast cancer tumor board decision making. Journal of Personalized Medicine 13:1502. https://doi.org/10.3390/jpm13101502

[5] Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG (2023) Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR)" Case of the Month". Cureus 15. https://doi.org/10.7759/cureus.43958

[6] Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, Weber T, Wesp P, Sabel BO, Ricke J (2023) ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol:1-9. https://doi.org/10.1007/s00330-023

[7] Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH (2024) Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis. Cardiovasc Intervent Radiol 47:245-250. https://doi.org/10.1007/s00270-023-03563-2

[8] Elkassem AA, Smith AD (2023) Potential use cases for ChatGPT in radiology reporting. American Journal of Roentgenology 221:373-376. https://doi.org/10.2214/AJR.23.29198

[9] Chan V, Perlas A (2011) Basics of ultrasound imaging. Atlas of ultrasound-guided procedures in interventional pain management:13-19. https://doi.org/10.1007/978-1-4419-1681-5_2

[10] Barratt A, Copp T, McCaffery K, Moynihan R, Nickel B (2017) Words do matter: a systematic review on how different terminology for the same condition influences management preferences. https://doi.org/10.1136/bmjopen-2016-014129

[11] Johnson AJ, Frankel RM, Williams LS, Glover S, Easterling D (2010) Patient access to radiology reports: what do physicians think? Journal of the American College of Radiology 7:281-289. https://doi.org/10.1016/j.jacr.2009.10.011

[12] Amin K, Khosla P, Doshi R, Chheang S, Forman HP (2023)

Focus: Big Data: Artificial Intelligence to Improve Patient Understanding of Radiology Reports. The Yale Journal of Biology and Medicine 96:407. https://doi.org/10.59249/NKOY5498

[13] Ateşman E (1997) Türkçede okunabilirliğin ölçülmesi. Dil Dergisi 58

[14] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, De Vet HC (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology 277:826-832. https://doi.org/10.1136/bmj.h5527

[15] Khan R, Gupta N, Sinhababu A, Chakravarty R (2023) Impact of Conversational and Generative AI Systems on Libraries: A Use Case Large Language Model (LLM). Science & Technology Libraries:1-15. https://doi.org/10.1080/0194262x.2023.2254814

[16] Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP (2024) Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis. Radiology 310:e231593. https://doi.org/10.1148/radiol.231593

[17] Haver HL, Gupta AK, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH (2024) Evaluating the Use of ChatGPT to Accurately Simplify Patient-centered Information about Breast Cancer Prevention and Screening. Radiology: Imaging Cancer 6:e230086. https://doi.org/10.1148/rycan.230086

[18] Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M (2023) Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. Digital Health 9:20552076231221620. https://doi.org/10.1177/20552076231221620

[19] Li H, Moon JT, Iyer D, Balthazar P, Krupinski EA, Bercu ZL, Newsome JM, Banerjee I, Gichoya JW, Trivedi HM (2023) Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. Clin Imaging 101:137-141. https://doi.org/10.1016/j.clinimag.2023.06.008

[20] Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, Wang G, Whitlow CT (2023) Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Visual Computing for Industry, Biomedicine, and Art 6:9. https://doi.org/10.1186/s42492-023-00136-5

[21] Tepe M, Emekli E (2024) Decoding medical jargon: The use of AI language models (ChatGPT-4, BARD, microsoft copilot) in radiology reports. Patient Educ Couns:108307. https://doi.org/10.1016/j.pec.2024.108307

How to Cite;

Gunes YC, Cesur T, Camur E (2024) Comparative Analysis of Large Language Models in Simplifying Turkish Ultrasound Reports to Enhance Patient Understanding. Eur J Ther. 30(5):714-723. https://doi.org/10.58600/eurjther2225