ISTANBUL UNIVERSITY PRESS

RESEARCH / ARAŞTIRMA
DOI: 10.26650/IUITFD.1494572
İst Tıp Fak Derg 2024 / J Ist Faculty Med 2024

A COMPARATIVE STUDY: PERFORMANCE OF LARGE LANGUAGE MODELS IN SIMPLIFYING TURKISH COMPUTED TOMOGRAPHY REPORTS

KARŞILAŞTIRMALI BİR ÇALIŞMA: TÜRKÇE BİLGİSAYARLI TOMOGRAFİ RAPORLARININ SADELEŞTİRİLMESİNDE BÜYÜK DİL MODELLERİNİN PERFORMANSI

Eren ÇAMUR[1], Turay CESUR[2], Yasin Celal GÜNEŞ[3]

[1]Department of Radiology, Ministry of Health Ankara 29 Mayis State Hospital, Ankara, Türkiye
[2]Department of Radiology, Ministry of Health Mamak State Hospital, Ankara, Türkiye
[3]Department of Radiology, Ministry of Health Kırıkkale Yüksek İhtisas Hospital, Kırıkkale, Türkiye

ORCID IDs of the authors: E.Ç. 0000-0002-8774-5800; T.C. 0000-0002-2726-8045; Y.C.G. 0000-0001-7631-854X

Cite this article as: Çamur E, Cesur T, Güneş YC. A comparative study: Performance of large language models in simplifying Turkish computed tomography reports. J Ist Faculty Med 2024;87(4):321-326. doi: 10.26650/IUITFD.1494572

ABSTRACT

Objective: This study evaluated the effectiveness of various large language models (LLMs) in simplifying Turkish Computed Tomography (CT) reports, a common imaging modality.

Material and Method: Using fictional CT findings, we followed the Standards for Reporting of Diagnostic Accuracy Studies (STARD) and the Declaration of Helsinki. Fifty fictional Turkish CT findings were generated. Four LLMs (ChatGPT 4, ChatGPT-3.5, Gemini 1.5 Pro, and Claude 3 Opus) simplified reports using the prompt: "Please explain them in a way that someone without a medical background can understand in Turkish." Evaluations were based on the Ateşman's Readability Index and Likert scale for accuracy and readability.

Results: Claude 3 Opus scored the highest in readability (58.9), followed by ChatGPT-3.5 (54.5), Gemini 1.5 Pro (53.7), and ChatGPT 4 (45.1). Likert scores for Claude 3 Opus (mean: 4.7) and ChatGPT 4 (mean: 4.5) showed no significant difference ($p>0.05$). ChatGPT 4 had the highest word count (96.98) compared to Claude 3 Opus (90.6), Gemini 1.5 Pro (74.4), and ChatGPT-3.5 (38.7) ($p<0.001$).

Conclusion: This study shows that LLMs can simplify Turkish CT reports at a level that individuals without medical knowledge can understand and with high readability and accuracy. ChatGPT 4 and Claude 3 Opus produced the most comprehensive simplifications. Claude 3 Opus' simpler sentences may make it the optimal choice for simplifying Turkish CT reports.

Keywords: Large language model, radiology reports, readability, computed tomography, Turkish, simplifying

ÖZET

Amaç: Bu çalışmada, yaygın bir görüntüleme yöntemi olan Türkçe bilgisayarlı tomografi (BT) raporlarının sadeleştirilmesinde çeşitli büyük dil modellerinin (BDM) etkinliği değerlendirilmiştir.

Gereç ve Yöntem: Kurgusal BT bulguları kullanılarak, Tanısal Doğruluk Çalışmaları Raporlama Standartları (STARD) ve Helsinki Bildirgesi'ne uyulmuştur. Elli kurgusal Türkçe BT bulgusu oluşturuldu. Dört LLM (ChatGPT 4, ChatGPT-3.5, Gemini 1.5 Pro ve Claude 3 Opus) istemini kullanarak raporları sadeleştirdi: "Please explain them in a way that someone without a medical background can understand in Turkish". Okunabilirlik değerlendirmesi Ateşman Okunabilirlik Endeksi, doğruluk derecesi Likert ölçeğine göre yapılmıştır.

Bulgular: Claude 3 Opus okunabilirlik açısından en yüksek puanı alırken (58,9), onu ChatGPT-3.5 (54,5), Gemini 1.5 Pro (53,7) ve ChatGPT 4 (45,1) izledi. Claude 3 Opus (ortalama: 4,7) ve ChatGPT 4 (ortalama: 4,5) için Likert skorları anlamlı bir farklılık yoktu ($p>0,05$). ChatGPT 4, Claude 3 Opus (90,6), Gemini 1.5 Pro (74,4) ve ChatGPT-3.5 (38,7) ile karşılaştırıldığında en yüksek kelime sayısına (96,98) sahipti ($p<0,001$).

Sonuç: Bu çalışma, BDM'lerin Türkçe BT raporlarını tıp bilgisi olmayan bireylerin anlayabileceği düzeyde ve yüksek okunabilirlik ve doğrulukla sadeleştirebildiğini göstermektedir. ChatGPT 4 ve Claude 3 Opus en doğru sadeleştirmeleri yapmaktadır. ChatGPT 4'ün daha basit cümleleri, onu Türkçe BT raporları için tercih edilen seçenek haline getirebilir.

Anahtar Kelimeler: Büyük dil modelleri, radyoloji raporları, okunabilirlik, bilgisayarlı tomografi, Türkçe, sadeleştirme

Corresponding author/İletişim kurulacak yazar: Eren ÇAMUR – eren.camur@outlook.com

Submitted/Başvuru: 03.06.2024 • **Revision Requested/Revizyon Talebi:** 19.08.2024 •
Last Revision Received/Son Revizyon: 27.08.2024 • **Accepted/Kabul:** 02.09.2024 • **Published Online/Online Yayın:** 07.10.2024

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

INTRODUCTION

Large language models (LLMs) have received considerable global attention, with numerous studies conducted worldwide. This is due to the sophisticated human-like communication and reasoning capabilities of these models (1, 2). As in many other matters, the performance of LLMs in radiological assessments, their familiarity with radiological guidelines, and their role in aiding differential diagnosis and occasionally making final decisions have recently attracted significant attention within the radiology community (3, 4).

Radiology reports, which provide a summary of radiologists' considerations and insights derived from imaging studies, are important in guiding diagnosis and treatment. They play a pivotal role in clinical practise, facilitating communication between healthcare providers and between patients and physicians. The capacity of LLMs to summarise, identify the principal ideas in texts, and interpret them has prompted an increasing interest in their potential to facilitate the simplification of radiology reports (5-8). This application would enable LLMs to enhance patients' understanding of radiology reports, alleviate their anxiety, and improve communication among healthcare providers and between patients and physicians.

This study compared how effectively different LLMs simplify Turkish CT reports, an imaging modality frequently obtained in clinical practise.

MATERIAL AND METHODS

The study included only fictional CT findings, excluding actual radiology reports, thereby negating the need for ethical board approval. The study design adhered to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) and the principles outlined in the Declaration of Helsinki (9).

The authors collaboratively generated 50 fictional CT findings in Turkish used in radiology reports. Efforts were made to ensure that these findings were representative of common scenarios in daily practise and depicted realistically (Supplementary material 1) (Table 1).

The study employed various LLMs, including ChatGPT 4, ChatGPT-3.5, Gemini 1.5 Pro, and Claude 3 Opus. The fictional findings were input into each LLM via their respective websites using the prompt, "I will write the findings from the CT report below. Please explain them in a way that someone without a medical background can understand" in Turkish (Figure 1a, 1b). Each finding was processed in a new window, as shown in Figure 1, with the default settings applied for each model. The study was conducted between April 15 and April 19, 2024.

The responses from the LLMs were evaluated using the Ateşman's Readability Index [198,825-(40,175xnumber of syllables/number of words)-(2,610xnumber of words/number of sentences)] to determine the readability levels (Table 2) (10). This analysis was performed using the publicly accessible and free website "www.readabilityindex.com." The three authors collectively evaluated the responses on a five-point Likert scale, with one representing the least favourable and five representing the most favourable, in terms of medical accuracy, consistency of recommendations, and comprehensibility. Additionally, the word count for each response was documented. The study's workflow is illustrated in Figure 2.

For the statistical analyses, we used SPSS ver.26 (IBM Corp, Armonk, NY, USA). Data distribution was assessed using the Kolmogorov-Smirnov and Shapiro-Wilk tests, while the Levene test was used to evaluate data variance. Descriptive statistics included the minimum, maximum, average, median, standard deviation, interquartile range, and percentages. To identify significant relationships between the quantitative data independent groups, we employed the Friedman and Wilcoxon tests. Spearman correlation analysis was used to examine the linearity of the correlations between the quantitative data.

RESULTS

There was no statistically significant difference between the Likert scores of Claude 3 Opus (mean: 4.7; median: 5.0) and ChatGPT 4 (mean: 4.5; median: 5.0) ($p>0.05$). However, Claude 3 Opus's Likert scores differed significantly from those of Gemini 1.5 Pro (mean: 4.3; median: 4.0) and ChatGPT-3.5 (mean: 2.8; median: 3.0) ($p<0.001$). While there was no statistically significant difference between the average Likert scores of ChatGPT 4 and Gemini 1.5 Pro ($p=0.025$; Bonferroni-adjusted p-value=0.0125), a significant difference was observed between the scores of ChatGPT 4 and ChatGPT-3.5 (mean: 2.8; median: 3.0) ($p<0.001$). The average Likert score for ChatGPT-3.5 was significantly lower than that of all other large language models ($p<0.001$).

According to Ateşman's readability index and readability levels, Claude 3 Opus had the highest average value at 58.9, followed by ChatGPT-3.5 (54.5), Gemini 1.5 Pro (53.7), and ChatGPT 4 (45.1). Although there was a significant difference in readability between Claude 3 Opus and ChatGPT 4 ($p<0.05$), there wasn't significant difference observed with other LLMs ($p>0.05$). The descriptive findings of the study are shown in Table 3.

A statistically significant difference was found in the number of words used by ChatGPT 4 (mean: 96.98) compared to Claude 3 Opus (mean: 90.6), Gemini 1.5 Pro (mean: 74.4), and ChatGPT-3.5 (mean: 38.7) ($p<0.001$). Although there was no significant difference in the word count between ChatGPT 4 and Claude 3 Opus, the average word

Table 1: A portion of the findings used as fictional Turkish and English CT findings are shown*.

1. Sol frontalde en kalın yerinde 2 mm ölçülen subaraknoid kanama izlendi (A subarachnoid haemorrhage, measuring 2 mm at its thickest point, was observed in the left frontal region)
2. Sol frontal lob komşuluğunda en kalın yerinde 20 mm ölçülen epidural kanama izlendi (An epidural haemorrhage, measuring 20 mm at its thickest point, was observed adjacent to the left frontal lobe)
3. Sol frontal lob komşuluğunda en kalın yerinde 30 mm ölçülen subdural kanama izlendi (A subdural haemorrhage, measuring 30 mm at its thickest point, was observed adjacent to the left frontal lobe)
4. Her iki frontal lobda atrofiye ikincil hemisferik kortikal sulkuslarda belirginleşme derinleşme izlendi (Enlargement and deepening of the hemispheric cortical sulci, secondary to atrophy, were observed in both frontal lobes)
5. Sağ temporal kemikte transvers fraktür izlendi (Transverse fracture was observed at the right temporal bone)
6. Sol plevral aralıkta en kalın yerinde 20 mm ölçülen plevral effüzyon izlendi (Pleural effusion, measuring 20 mm at its thickest point, was observed in the left pleural space)
7. Sağ akciğer alt lobda konsolidasyon tarzında infiltrasyon izlendi (Consolidation-infiltration was observed in the lower lobe of the right lung)
8. Her iki akciğer apekste sekel fibrotik değişiklikler izlendi (Sequelae fibrotic changes were observed at the apices of both lungs)
9. Sol akciğer lingüler segmentte atelektatik değişiklikler izlendi (Atelectatic changes were observed in the lingular segment of the left lung)
10. Perikardial aralıkta en kalın yerinde 11 mm ölçülen perikardial effüzyon izlendi (Pericardial effusion, measuring 11 mm at its thickest point, was observed in the pericardial space)
11. Kardiyotorasik oran kalp lehine artmıştır (The cardiothoracic ratio is increased in favour of the heart)
12. Pulmoner trunk 37 mm ölçülmüş olup ektatiktir (The pulmonary trunk was measured at 37 mm, indicating ectasia)
13. Sol akciğer alt lobda 7 mm çapında solid nodül izlendi (A solid nodule, measuring 7 mm in diameter, was observed in the lower lobe of the left lung)
14. Sol akciğer alt lobda 7 mm çapında semi-solid nodül izlendi (A semi-solid nodule, measuring 7 mm in diameter, was observed in the lower lobe of the left lung)
15. Göğüs on-arka çapı belirgin artmıştır (The anteroposterior diameter of the chest is markedly increased)
16. Tiroid gland boyutları belirgin artmış olup trakea sola itilmiştir (The dimensions of the thyroid gland are significantly increased, with the trachea displaced to the left)
17. Karaciğer parankiminde steatoza ikincil diffüz dansite azalması izlendi (Diffuse decrease in parenchymal density, secondary to steatosis, was observed in the liver)
18. Karaciğerde 10mm çapında hemanjiyom ile uyumlu periferik nodüler kontrastlanana hipodens lezyon izlendi (A hypodense lesion, measuring 10 mm in diameter and consistent with a hemangioma, exhibiting peripheral nodular enhancement, was observed in the liver)
19. Karaciğerde 15mm çapında kontrastlanmayan öncelikle basit kist lehine düşünülen hipodens lezyon izlendi (A hypodense lesion, measuring 15 mm in diameter and not enhancing with contrast, was observed in the liver, favouring a diagnosis of a simple cyst)
20. Safra kesesi fundus düzeyinde fokal duvar kalınlık artışı izlendi (Focal wall thickening was observed at the fundus of the gallbladder)

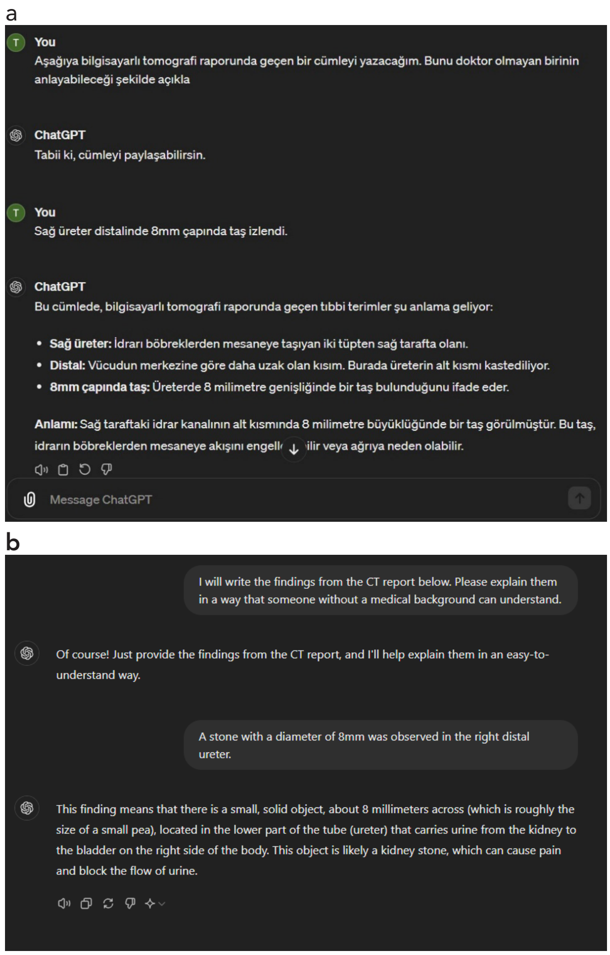
*:Since the findings in the study are given to LLMs in Turkish, the findings are presented both in Turkish and in brackets in English.

count for ChatGPT 4 and Claude 3 Opus was statistically higher than that of all other language models (p<0.001). Additionally, the word count for Gemini 1.5 Pro was statistically higher than that for ChatGPT-3.5 (p<0.001).

A linear correlation was observed between the number of words in the fictional CT findings and those generated by Gemini 1.5 Pro (correlation coefficient=0.756, p<0.000) and ChatGPT 4 (correlation coefficient=0.523, p<0.000). In contrast, no linear correlation was detected for Claude 3 Opus (p=0.367) and Perplexity (p=0.552). Additionally, a linear correlation was identified between the readability index of the fictional CT findings and the readability indices of the responses from Claude 3 Opus (correlation coefficient=0.508, p<0.000), ChatGPT 3,5 (correlation coefficient=0.387, p=0.005), and Gemini 1.5 Pro (correlation coefficient=0.331, p=0.042). However, no correlation was found between the readability index of the fictional CT findings and that of ChatGPT 4 (p=0.387).

DISCUSSION

The main result of our study is that LLMs can simplify CT reports with high readability and accuracy. The Likert scale evaluations by all three radiologists indicated a consensus that LLMs can accurately simplify CT reports in a medically sound manner. Similarly, previous studies have noted that LLMs are highly effective in simplifying and explaining radiology reports (5, 6, 10, 11). Jeblick et al. evaluated the quality of ChatGPT's simplified English radiology reports, including fictional reports, and found them to be accurate and complete, based on medical



You
Aşağıya bilgisayarlı tomografi raporunda geçen bir cümleyi yazacağım. Bunu doktor olmayan birinin anlayabileceği şekilde açıkla

ChatGPT
Tabii ki, cümleyi paylaşabilirsin.

You
Sağ üreter distalinde 8mm çapında taş izlendi.

ChatGPT
Bu cümlede, bilgisayarlı tomografi raporunda geçen tıbbi terimler şu anlama geliyor:

- **Sağ üreter:** İdrarı böbreklerden mesaneye taşıyan iki tüpten sağ tarafta olanı.
- **Distal:** Vücudun merkezine göre daha uzak olan kısım. Burada üreterin alt kısmı kastediliyor.
- **8mm çapında taş:** Üreterde 8 milimetre genişliğinde bir taş bulunduğunu ifade eder.

Anlamı: Sağ taraftaki idrar kanalının alt kısmında 8 milimetre büyüklüğünde bir taş görülmüştür. Bu taş, idrarın böbreklerden mesaneye akışını engell ilir veya ağrıya neden olabilir.

I will write the findings from the CT report below. Please explain them in a way that someone without a medical background can understand.

Of course! Just provide the findings from the CT report, and I'll help explain them in an easy-to-understand way.

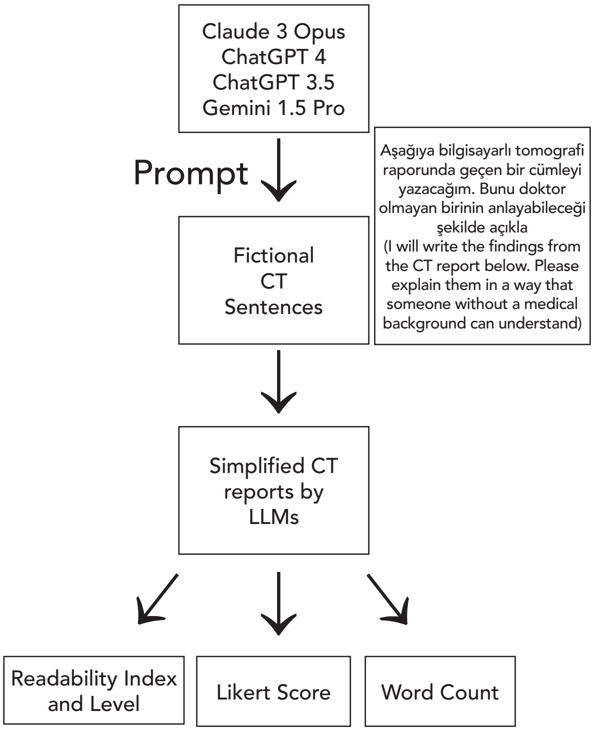
A stone with a diameter of 8mm was observed in the right distal ureter.

This finding means that there is a small, solid object, about 8 millimeters across (which is roughly the size of a small pea), located in the lower part of the tube (ureter) that carries urine from the kidney to the bladder on the right side of the body. This object is likely a kidney stone, which can cause pain and block the flow of urine.

Figure 1: a) Describes the Turkish input and output process in the study through ChatGPT 4. The figure shows the Turkish version of the prompt mentioned in the methodology and the fictional finding as input. **b)** Describes the English input and output process in the study through ChatGPT 4. The figure shows the English version of the prompt mentioned in the methodology and the fictional finding as input. Explanation: Since the findings of the study are given to LLMs in Turkish, the findings are presented in Turkish in Figure 1a and in English in Figure 1b.

facts, suggesting that ChatGPT can achieve this simplification without causing any harm to patients (10).

We used Ateşman's readability index to assess how easily the simplified CT reports, produced by LLMs in Turkish, could be read (12). It measures Turkish text readability based on average syllables per word and words per sentence, with scores ranging from 1 to 100; higher scores indicate easier reading. Ateşman stressed that a text's effectiveness relies on both its readability and comprehensibility. While readability is quantitatively evaluated, comprehensibility is qualitatively assessed



Aşağıya bilgisayarlı tomografi raporunda geçen bir cümleyi yazacağım. Bunu doktor olmayan birinin anlayabileceği şekilde açıkla
(I will write the findings from the CT report below. Please explain them in a way that someone without a medical background can understand)

Figure 2: The study's workflow.
The prompt in the study was given to LLMs in Turkish. Thus, the prompt is presented both in Turkish and in brackets in English.

Table 2: Ateşman's Readability Index and its corresponding readability level

Index	Readability level
90-100	Easily understood by 4thgrade 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 study are shown.

	Claude 3 Opus	Gemini 1.5pro	ChatGPT 4	ChatGPT 3.5
Likert Scores*				
Minimum-Maximum	4.0-5.0	3.0-5.0	3.0-5.0	1.0-4.0
Mean±SD	4.7±0.4	4.32±0.6	4.52±0.5	2.78±0.6
Median (IQR)	5.0 (0)	4.0 (0)	5.0 (0)	3.0 (1.0)
Ateşman's Readability Index				
Minimum-Maximum	33.1-79.0	23.3-68.9	34.9-72.3	21.8–79.1
Mean±SD	58.9±5.23	53.7±5.67	45.1±11.72	54.5±9.99
Median (IQR)	53.4 (13.2)	48.2 (10.9)	38.1 (8.2)	38.3 (7.1)
Readability Level				
Minimum	7-8th class	9-10th class	7-8th class	7-8th class
Maximum	Bachelor's degree	Postgraduate	Bachelor's degree	Postgraduate
Median	11-12th class	11-12th class	13-14th class	11-12th class
Word Count				
Minimum-Maximum	71-136	34-143	47-149	8-71
Mean±SD	90.66±16.77	74.42±27.26	96.98±28.39	38.74±15.9
Median (IQR)	85.0 (24.5)	69.0 (43.7)	97.5 (34.5)	40.0 (30.0)

*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. SD: Standard Deviation, IQR: Interquartile range.

based on the text's content. We evaluated the readability of the responses using the Ateşman's index and their comprehensibility using a Likert scale. We acknowledge that assessments by individuals without medical backgrounds would offer more valuable insights into comprehensibility. There was no significant difference between ChatGPT 4 and Claude 3 Opus in terms of Likert score, but the readability index of ChatGPT 4 was lower than all other LLMs. Claude 3 Opus had both the highest Likert score and the highest Ateşman's index among all LLMs. This shows that Claude 3 Opus uses more simple and understandable sentences to simplify CT reports by providing sufficient and accurate information. Claude 3 Opus adeptly simplifies Turkish CT reports while employing straightforward sentence structures. Hence, Claude 3 Opus may represent the optimal choice among LLMs for streamlining Turkish CT reports.

Johnson et al. simplified 750 randomly selected anonymized radiology reports with three different prompts (x-ray, ultrasound, magnetic resonance, and computed tomography reports) using ChatGPT 3.5, ChatGPT 4, Microsoft Bing, and Google Bard (now known as Gemini). The researchers reported that all LLMs produced more readable reports than the original reports (13). They also reported that the performance of each LLM was affected differently at different prompts. Although there is no generally accepted prompt for report simplification, the prompt given significantly influences LLM responses. Lyu et al. examined 62 thorax CT and 76 brain MRI reports (14). Each report had three simplified versions based on different prompts: making the report easier to understand, providing patient advice, and offering healthcare professional recommendations. They also explored how different prompts could create varied reports for patients with different education levels. Similarly, Schmidt et al. used ChatGPT 3.5 to simplify knee MRI findings of varying complexity (simple, moderate, and complex) with five different prompts (11). They showed the effect of prompts on simplifying radiology reports. In addition, their findings revealed that simplified reports were more comprehensible for patients, leading to improved patient understanding and overall satisfaction. Li et al. simplified 100 radiology reports, including different imaging modalities, using the prompt "Explain this radiology report to a patient in layman's terms: <Report Text>" and showed that simplified reports were significantly more readable and shorter (6). In order not to affect the Likert scores and readability levels of each LLM, we were careful not to include specific words that might affect the word limit and readability level of our prompt. Further studies will be instructive to show how the specific prompts given affect the readability level. In this way, the information content and readability level of the simplified reports can be adjusted by giving specific prompts according to the socio-economic level of the patients.

Our study is the first to assess how LLMs can simplify Turkish CT reports to understand people without a medical background. However, it has some limitations. The main limitation is that only radiologists scored simplified reports. Practitioners from other departments and real patients did not participate in this study, so we lacked their feedback on the simplified CT reports. Future research should include patient feedback and compare standard CT reports with those simplified by LLMs. This would provide important insights into how understandable and useful the simplified reports are to patients. In addition, we only used fictitious findings for a single condition, not real CT reports. More complex reports that included all relevant findings might produce different results. Finally, we used only one prompt. Different prompts could produce better or worse results depending on the capabilities of the model.

CONCLUSSION

In conclusion, our study shows that LLMs can effectively simplify Turkish CT reports. Enabling patients to read and understand CT reports may help them better grasp their diagnosis and treatment, leading to improved compliance. Simplified CT reports may also enhance communication between physicians.

Ethics Committee Approval: The study does not require ethics committee approval because it only includes fictional computed tomography findings and does not use actual radiology reports.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- E.Ç., T.C.; Data Acquisition – E.Ç., T.C., Y.C.G.; Data Analysis/Interpretation- E.Ç.; Drafting Manuscript- E.Ç.; Critical Revision of Manuscript- T.C., Y.C.G.; Final Approval and Accountability- E.Ç.; Supervision- T.C., Y.C.G.

Conflict of Interest: The authors have no conflict of interest to declare.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A Survey of Large Language Models. 2023 http://arxiv.org/abs/2303.18223
2. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon L De, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health 2023;2(2):e0000198. [CrossRef]
3. Yilmaz EC, Belue MJ, Turkbey B, Reinhold C, Choyke PL. A Brief Review of Artificial Intelligence in Genitourinary Oncological Imaging. Can Assoc Radiol J 2023;74(3):534-47. [CrossRef]
4. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagnostic and Interventional Radiology 2024;30(2):80-90. [CrossRef]
5. Doshi R, Amin K, Khosla P, Bajaj S, Chheang S, Forman HP. Utilizing Large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing. medRxiv 2023. https://www.medrxiv.org/content/10.1101/2023.06.04.23290786v2 [CrossRef]
6. Li H, Moon JT, Iyer D, Balthazar P, Krupinski EA, Bercu ZL, et al. Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. Clin Imaging. 2023;101:137-41. [CrossRef]
7. Luo W, Liu F, Liu Z, Litman D. A novel ILP framework for summarizing content with high lexical variety. Nat Lang Eng 2018;24(6):887-920. [CrossRef]
8. Guadalupe Ramos J, Navarro-Alatorre I, Flores Becerra G, Flores-Sánchez O. A Formal Technique for Text Summarization from Web Pages by using Latent Semantic Analysis. Research in Computing Science 2019;148(3):11-22. [CrossRef]
9. Bossuyt PM, Reitsma JB, Bruns DE, Bruns DE, Glasziou PP, Irwig L, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies1. Radiology 2015;277(3):826-32. [CrossRef]
10. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2023;1:1-9. [CrossRef]
11. Schmidt S, Zimmerer A, Cucos T, Feucht M, Navas L. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. Arch Orthop Trauma Surg 2024;144(2):611-8. [CrossRef]
12. Ateşman E. Türkçede okunabilirliğin ölçülmesi. Dil Dergisi. 1997;58:71-4.
13. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10(1):1. [CrossRef]
14. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6(1):1-10. [CrossRef]