Uludağ Üniversitesi Tıp Fakültesi Dergisi
Journal of Uludağ University Medical Faculty
51 (2) 305-309, 2025
DOI: https://doi.org/10.32708/uutfd.1653680

ORIGINAL RESEARCH

Evaluating the Performance of Large Language Models in Generating Impressions for Radiology Reports

Hasan Emin KAYA, Dilek SAĞLAM, Zeynep YAZICI, Gökhan GÖKALP

Bursa Uludağ University, Faculty of Medicine, Department of Radiology, Bursa, Türkiye.

ABSTRACT

The aim of the study was to evaluate and compare the performance of three popular large language models (LLMs) in generating impressions for radiology reports in Turkish. ChatGPT, Gemini, and Copilot were used to generate impressions for 50 anonymized radiology reports using a "few-shot" prompt. The impressions were scored by three radiologists using a Likert scale, based on whether they included all relevant information from the report, provided an appropriate summary of the report, contained no misleading information, and could be added to the report without modification. Friedman's test was used to evaluate whether there was a difference between the scores of the LLMs. The 50 reports included 32 magnetic resonance examinations, 11 computed tomography examinations, 5 ultrasound examinations, and 2 fluoroscopy examinations. Of these, 15 were neuroradiology studies, 14 were musculoskeletal studies, 13 were abdominal studies, and 8 were thoracic radiology studies. The median scores for the models' outputs were 4 and 5. This finding indicates that the radiologists generally found the models successful in generating impressions. Furthermore, no statistically significant difference was found among the models in terms of their performance in containing all information, providing an appropriate summary, avoiding misleading information, and being suitable for inclusion in the report without modification (p = 0.607, 0.327, 0.629, 0.089, respectively). In conclusion, ChatGPT, Gemini, and Copilot were found to be successful in generating impressions for radiology reports in Turkish, and no significant difference in performance was detected among the models.

Keywords: Radiology. Artificial intelligence. Large language models.

Büyük Dil Modellerinin Radyoloji Raporları İçin Sonuç Bölümü Oluşturmadaki Performanslarının Değerlendirilmesi

ÖZET

Çalışmamızın amacı popüler üç büyük dil modelinin (BDM) Türkçe radyoloji raporları için sonuç bölümü oluşturma konusundaki performansını değerlendirip mukayese etmekti. Anonimize edilmiş 50 radyoloji raporu için, "few-shot" bir komut ile, ChatGPT, Gemini ve Copilot dil modellerine sonuç bölümü oluşturuldu. Sonuçlar; rapordaki tüm bilgileri içerme, raporu uygun bir şekilde özetleme, yanıltıcı bilgi içermeme ve değiştirilmeden rapora eklenebilme açısından üç radyolog tarafından bir Likert skalası kullanılarak skorlandı. Friedman testi ile BDM'lerin skorları arasında fark olup olmadığı değerlendirildi. Çalışmaya dahil edilen 50 raporun 32'si manyetik rezonans, 11'i bilgisayarlı tomografi, 5'i ultrason ve 2'si floroskopi tetkikleriydi. Bu tetkiklerden 15'i nöroradyoloji, 14'ü kas-iskelet, 13'ü abdomen ve 8'i toraks radyolojisi çalışmalarıydı. Üç radyoloğun yaptığı skorlamalarda modellerin aldığı skorların medyan değerleri 4 ve 5 idi. Bu bulgu modellerin sonuç oluşturmada radyologlar tarafından genel olarak başarılı bulunduğunu göstermekteydi. Ayrıca modeller arasında bütün bilgileri içerme, raporu uygun bir şekilde özetleme, yanıltıcı bilgi içermeme ve değiştirilmeden rapora eklenebilme performansı açısından istatistiksel bir farklılık saptanmadı (*p* değerleri sırasıyla 0,607; 0,327; 0,629; 0,089). Sonuç olarak ChatGPT, Gemini ve Copilot Türkçe radyoloji raporları için sonuç bölümü oluşturmada başarılı bulunmuş ve modellerin performansı arasında anlamlı bir farklılık saptanmamıştır.

Anahtar Kelimeler: Radyoloji. Yapay zeka. Büyük dil modelleri.

Large language models (LMMs) are advanced artificial intelligence applications built using deep neural networks and trained using very large amounts of text to generate human-like responses. ChatGPT, one of the most popular of these models, is developed by OpenAI and designed as a chatbot that can interact with users through deep learning and natural language processing algorithms. Gemini, developed by Google, is another artificial intelligence model that can engage in text-based interactions with humans thanks to similar natural language processing capabilities. Copilot, developed by Microsoft, is another popular artificial intelligence model built using natural language processing and deep learning algorithms.

Date Received: 10.March.2025
Date Accepted: 31.July.2025

AUTHORS' ORCID INFORMATION
Hasan Emin KAYA: 0000-0002-7411-4102
Dilek SAĞLAM: 0000-0002-5778-6847
Zeynep YAZICI: 0000-0002-8647-5298
Gökhan GÖKALP: 0000-0002-3682-2474

It has been suggested that ChatGPT can be useful for radiologists in reporting by assisting in creating clinical information or impression sections, summarizing reports for patients, and increasing patient interaction[1]. In a study evaluating the performance of ChatGPT, Google Bard (now Gemini), and Microsoft Bing (now Copilot) in simplifying radiology reports, the models were shown to be able to perform this task accurately[2]. In another study evaluating ChatGPT's performance in generating impression sections for radiology reports, the impressions produced by the model scored lower than those generated by radiologists[3]. A language model specifically developed to generate radiology report impressions has successfully produced professional and linguistically appropriate impressions for a wide range of radiological examinations[4].

In studies evaluating the performance of popular LLMs in generating impressions for radiology reports, the models have generally been used with "zero-shot" prompts. "Zero-shot" refers to a model's ability to perform a new task without having been explicitly trained on examples of that specific task. On the other hand, "few-shot" learning in LLMs refers to the ability of the model to perform a task after being provided with a few examples (or "shots"). The aim of our study is to use three popular LLMs (ChatGPT, Gemini, and Copilot) to generate impressions for radiology reports in Turkish using a few-shot prompt, evaluate the appropriateness of these impressions, and compare the performance of the LLMs in this task.

Material and Method

After obtaining approval for the study from the ethics committee of our university (Decision number: 2024-19/1), 50 radiology reports from our institution's picture archiving and communication system (PACS) that were created in 2024 were selected and anonymized. The impressions of these reports were then removed, leaving only the clinical information section and the body of the report. Then, a detailed prompt was prepared for the language models (ChatGPT o1, Gemini 1.5 Pro, Copilot) to create an impression for the reports. When creating the prompt, instead of a "zero-shot" prompt such as "simplify this report"[5], a more detailed prompt was written to obtain an impression more similar to what radiologists create during their daily practice. Additionally, two examples were given to make it easier for the models to learn to do the desired task. Below is the English translation of the prompt used in the study (Original Turkish prompt can be found as a supplementary file):

"You are a radiologist. Create an impression for the radiology report provided. When doing this, follow these guidelines: 1) Create the impression in a way that appeals to health professionals. 2) Do not include normal structures and findings in the impression. 3) Include only pathologic findings in the impression. 4) Specify the diagnosis or possible diagnoses by interpreting the findings. 5) Make the impression as concise as possible.

I have given an example below.

Clinical Information: Knee pain

Findings: The amount of fluid within the knee joint is increased. There is a horizontal tear in the posterior horn of the medial meniscus. No tear in the lateral meniscus. Anterior and posterior cruciate ligaments are intact. Medial and lateral collateral ligaments are intact. Quadriceps and patellar tendons are normal. Cartilage defects involving less than 50% thickness are observed in the medial femorotibial joint. Subchondral bone marrow edema-like signal intensity is noted on posterior medial tibial plateau. No bone or soft tissue masses.

Impression: 1. Effusion 2. Horizontal tear of the posterior horn of the medial meniscus 3. Cartilage defects involving less than 50% thickness in the medial femorotibial joint 4. Subchondral bone marrow edema-like signal intensity on posterior medial tibial plateau.

Another example:

Clinical Information: Liver mass characterization

Findings: A mass lesion of approximately 5.5 cm in diameter is observed on segment 8 of the liver. There is a T2 hyperintense branching structure in the center of the lesion (consistent with vascular scar). The mass is slightly hypointense on T1-weighted images and iso-hyperintense on T2-weighted images compared to the liver parenchyma. On dynamic examination, the lesion enhances in the arterial phase and does not show washout. In the hepatobiliary phase, the enhancement is heterogeneous but persistent. No lesion in other parts of the liver. Gallbladder and bile ducts are normal. Spleen, pancreas, both adrenal glands, and kidneys are normal. No intra-abdominal free fluid, localized fluid collections, or lymphadenopathy are observed. The bladder and bony structures in the imaging field are normal.

Impression: Focal nodular hyperplasia in segment 8 of the liver.

Now generate an impression to this report in accordance with the above instructions and examples:"

To assess the quality and appropriateness of the impressions generated by the models, three radiologists with 27, 10, and 6 years of experience were asked to score the following four statements about the impressions: 1) The impression generated by the model contains all the necessary information specified in the report. 2) The impression generated by

the model summarizes the report appropriately. 3) The impression generated by the model does not contain false or misleading information. 4) I can add the impression generated by the model to my report without editing it. A 5-point Likert scale was used for scoring (1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly agree). A total of 12 scores were obtained for each impression (4 statements x 3 radiologists). Scores are reported as median and interquartile ranges. Friedman's test was used to assess differences among the LLM scores. Data analysis was performed using SPSS software (IBM SPSS Statistics for Windows, Version 27.0; Armonk, NY: IBM Corp.). A p-value < 0.05 was deemed significant.

Results

Of the 50 reports retrieved from PACS and anonymized, 32 were magnetic resonance imaging (MRI), 11 were computed tomography (CT), 5 were ultrasound and 2 were fluoroscopy studies. Of these, 15 were neuroradiology, 14 were musculoskeletal, 13 were abdominal and 8 were thoracic radiology examinations. The median scores provided by the three radiologists regarding whether the impressions generated by the models included all necessary information, summarized the reports appropriately, contained no misleading information, and could be added without modification were 4 and 5 (Table). This finding indicates that the models were generally considered successful by the radiologists in generating impressions (Figures 1a–d). In addition, no statistical difference was found between the models in terms of including all information, summarizing the report appropriately, not containing misleading information and being able to be added to the report without editing (Table).

Table. Likert scores of the LLMs (Data are given as median and interquartile range [25th and 75th percentile])

	ChatGPT	Gemini	Copilot	*p**
The impression generated by the model contains all the necessary information specified in the report.	5 (5–5)	5 (5–5)	5 (5–5)	0.607
The impression generated by the model summarizes the report appropriately.	5 (5–5)	5 (4–5)	5 (4–5)	0.327
The impression generated by the model does not contain false or misleading information.	5 (5–5)	5 (5–5)	5 (5–5)	0.629
I can add the impression generated by the model to my report without editing it.	5 (4–5)	5 (4–5)	4 (4–5)	0.089

*Friedman test

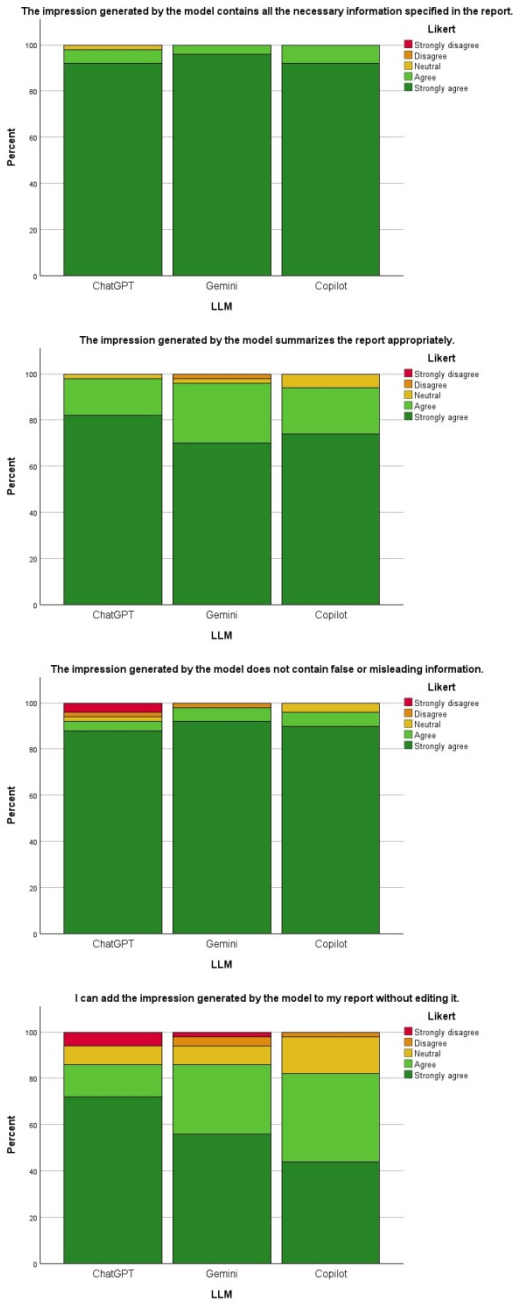

Figure 1.
Stacked bar charts of the scores of LLMs for (a) including all information, (b) summarizing the report appropriately, (c) not containing misleading information, and (d) being able to be added to the report without editing.

Discussion and Conclusion

Our study evaluating the performance of three popular LLMs in generating impressions for radiology reports in Turkish shows that the models can successfully perform this task and that there is no significant difference between the performance of the models.

Studies on the use of LLMs to generate impressions for radiology reports have focused more on evaluating the success of the models in simplifying radiology reports and generating patient-friendly impressions. For example, in Doshi et al.'s study evaluating the performance of ChatGPT-3.5, ChatGPT-4, Bard (now Gemini) and Bing (now Copilot) in generating simplified impressions for radiology reports[5] the authors used the following prompts: "Simplify this radiology report", "I am a patient. Simplify this radiology report", and "Simplify this radiology report at the 7th grade level". Can et al. used GPT-4, GPT-3.5 Turbo, Claude-3-Opus, Gemini Ultra as well as open-source models such as Mistral-7b and Mistral-8×7b to evaluate the performance of the models in simplifying interventional radiology reports in a way that patients can easily understand[6]. There are not many studies assessing the performance of the models in creating impressions that radiologists can include in their reports. In the study by Sun et al., the authors asked GPT-4 to *"*Generate a new short one-line impression from the findings section using medical vocabulary*"* for 50 chest x-rays[3]. However, in this study, the radiologist-generated impressions were shown to be better than the model-generated ones in terms of coherence, comprehensiveness, factual consistency and medical harmfulness. We believe that the better results we found in our study are related to the more detailed prompt we used and the fact that we used a few-shot prompt instead of a zero-shot one as used by Sun et al.

Zhang et al. have developed a new language model designed for creating impressions for radiology reports, using 20 gigabytes of medical and general purpose text for pre-training, and the impressions generated by the model were found to be in close agreement with the impressions of radiologists (median, 5 [IQR, 5–5] vs 5 [IQR, 5–5])[4]. This study shows that language models can be made more effective for specific tasks through fine-tuning.

Although the impressions generated by the models were generally appropriate, they were not entirely free of errors. For example, when presented with a chest CT showing unilateral absence of the pulmonary artery, ChatGPT suggested a potential diagnosis of scimitar syndrome. Likewise, for a pituitary MRI revealing a Rathke cleft cyst, the model's impression leaned toward a microadenoma. These cases illustrate that, in their current form, the models still require supervision and cannot be relied upon for independent use.

Our primary aim in this study was to evaluate how effectively the LLMs generate accurate, concise summaries suitable for daily practice, rather than assessing their diagnostic accuracy. Indeed, there are many studies in the literature evaluating the success of LLMs in diagnosing using text-based data. In a study evaluating the performance of ChatGPT on diagnosing cases published in the "Diagnosis Please" section of the journal Radiology, the accuracy of the model was 54%[7]. In another study evaluating the performance of ChatGPT in correctly diagnosing 100 "Case of the Week" examples published in the American Journal of Neuroradiology, the diagnostic accuracy of the model was found to be 50%[8]. In our study, none of the models could generate an impression section with the correct diagnosis for a brain MRI report with findings consistent with Aicardi syndrome. In an ankle MRI report of a tenosynovial giant cell tumor in the tibialis posterior tendon sheath, two of the models (ChatGPT and Gemini) were able to generate an impression with the correct diagnosis.

Limitations of our study include its retrospective nature and being a single-center study. Different reporting habits of different centers may affect the evaluation of the performance of the models. In addition, although we tried to create a detailed prompt, it may be possible to see the actual performance of the models by better prompting. In addition, since the models do not have access to the patients' electronic files, they lack some important laboratory or clinical information, which may also affect their performance. The impressions of radiology reports are written mostly for the referring clinician. The fact that we did not evaluate what clinicians think about the impressions generated by the models can be considered another limitation of this study.

Our study found that three popular LLMs generated generally appropriate impressions for radiology reports, with no significant differences in their performance. However, larger multi-center studies involving clinicians may provide a more comprehensive assessment of the effectiveness of LLMs in this task.

Researcher Contribution Statement:
Idea and design: H.E.K., D.S.; Data collection and processing: H.E.K., D.S., Z.Y.; Analysis and interpretation of data: H.E.K., D.S., Z.Y., G.G.; Writing of significant parts of the article: H.E.K., D.S.
Support and Acknowledgement Statement:
N/A
Conflict of Interest Statement:
The authors of the article have no conflict of interest declarations.
Ethics Committee Approval Information:
Approving Committee: Uludağ University Faculty of Medicine Clinical Research Ethics Committee
Approval Date: 20.11.2024
Decision No: 2024-19/1

References

1. Elkassem AA, Smith AD. Potential Use Cases for ChatGPT in Radiology Reporting. AJR Am J Roentgenol 2023;221(3):373–6.
2. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. Radiology 2023;309(2).

3. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J, et al. Evaluating GPT4 on Impressions Generation in Radiology Reports. Radiology 2023;307(5).
4. Zhang L, Liu M, Wang L, Zhang Y, Xu X, Pan Z, et al. Constructing a Large Language Model to Generate Impressions from Findings in Radiology Reports. Radiology 2024;312(3).
5. Doshi R, Amin KS, Khosla P, Bajaj S, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. Radiology 2024;310(3).
6. Can E, Uller W, Vogt K, Doppler MC, Busch F, Bayerl N, et al. Large Language Models for Simplified Interventional Radiology Reports: A Comparative Analysis. Acad Radiol. 2024 Sep 30:S1076-6332(24)00690-1. doi: 10.1016/j.acra.2024.09.041. Epub ahead of print. PMID: 39353826.
7. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. Radiology 2023;308(1).
8. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology 2024;66(1):73–9.