# ChatGPT-generated PICC Insertion Reports: A Simulation Study Using the RSNA Template

Jan F. Senge[1,2], Matthew T. McMurray[3], Fabian Haupt[3], Philipe S. Breiding[4], Claus Beisbart[5,6], Keivan Daneshvar[3], Alois Komarek[3], Gerd Nöldge[3], Frank Mosler[3], Wolfram A. Bosbach[3]

[1] Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany

[2] Max-Planck Dioscuri Centre for Topological Data Analysis, Warsaw, Poland

[3] Department of Diagnostic, Interventional and Pediatric Radiology (DIPR), Inselspital, Bern University Hospital, University of Bern, Switzerland

[4] University Institute of Diagnostic and Interventional Neuroradiology, Inselspital, Bern University Hospital, University of Bern, Switzerland

[5] Institute of Philosophy, University of Bern, Bern, Switzerland

[6] Center for Artificial Intelligence in Medicine, University of Bern, Bern, Switzerland

Corresponding author: Wolfram A. Bosbach · WolframAndreas.Bosbach@Insel.CH

Received: 4 April 2024 · Accepted: 24 April 2024 · Published: 27 April 2024

Swiss Journal of Radiology and Nuclear Medicine · https://www.sjoranm.com · SJORANM GmbH · Switzerland

## Abstract

**Motive:** Documentation and administration occupy a substantial part of the working time in interventional radiology. Automated drafting tools may assist with these tasks, but their potential benefit requires evaluation.

**Method:** Three predefined simulated cases of peripherally inserted central catheter (PICC) insertion were used to generate reports with ChatGPT on 4 May 2023, using items from the Radiological Society of North America (RSNA) template. Five iterations per case produced 15 report drafts. Eight radiologists, including six board-certified radiologists and two residents, rated completeness, structure, apparent AI authorship, readiness for use without editing and anticipated time savings.

**Key results:** Raters generally assessed completeness and structure positively and anticipated time savings. They generally disagreed that the drafts were clearly identifiable as AI-generated. Some drafts required editing, particularly in relation to report structure and the handling of complications. Documentation time was not measured, and a controlled comparison with human-written reports was not performed.

**Conclusions:** This small simulation study supports the feasibility of AI-assisted PICC report drafting and describes favourable perceptions among the participating raters. It does not establish actual time savings, equivalent report quality or indistinguishability from human-written reports. Prospective evaluation of drafting time, editing requirements and clinical accuracy is needed before conclusions about routine clinical benefit can be drawn.

Keywords: large language models; interventional radiology; automation of administration and documentation

## Introduction

In radiology, interventional radiology (IR) is the subspecialty which uses imaging for guiding minimally invasive surgical procedures. Imaging modalities applied today include fluoroscopy, ultrasound (US), computed tomography (CT), or magnetic resonance imaging (MRI) [1]. Our study investigated how IR could benefit from automated text drafting tools. We tested for the template of the Radiological Society of North America (RSNA) for peripherally inserted central catheter (PICC) lines [2] whether reports can be drafted by artificial intelligence (AI) based natural language processing models, ie ChatGPT [3].

Today, the interventional radiologist spends a substantial amount of time on administration / documentation work. As in other medical fields [4], [5], this activity is seen as an unpleasant necessity. It does not serve the immediate patient outcome. At the same time, demand for clinical radiology services is predicted to continue to grow in the future to a level that might not be able to be met by the workforce in its size today [6], [7]. This is why time savings through the use of AI text drafting would be a valuable and welcome contribution to the future of IR, from the viewpoint of patients, doctors, and hospital administrators alike.

Initial steps towards the computing technology required for AI in IR and elsewhere can be found in the work of Konrad Zuse [8]. The development of AI using digital computers was first proposed in eg [9], [10]. As other subfields of AI, natural language processing in combination with reinforcement learning, has recently seen some remarkable


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

advances [3], [11]. This is a broader development, not limited to the tool applied in the present study [12]. Regarding the application in radiology and elsewhere, the strengths of properly trained language processing lie in the huge knowledge base that is made available [13], and in the ability to communicate in different styles of language [14]. So far, rather few studies on AI based language processing have been published in IR. One relevant study has demonstrated limitations concerning accuracy of recommendations for IR procedures [15]. This result is similar to what was found in a previous study about the handling of technical and medical information in report drafting for distal radius fracture [16], [17].

For evaluating the ability of ChatGPT to handle the RSNA PICC line template [2], we defined three distinct simulated cases and iterated each case five times. Output texts were evaluated for content similarity and rated by eight radiologists. The main focus was to explore the feasibility of automated PICC report drafting and the participating radiologists' perceptions of report quality and potential time savings.

## Materials and Methods

The methodology of the presented study follows the concept of the previous work [16]: cases were defined within the framework of a current RSNA template. ChatGPT was tasked with report drafting. The output texts were evaluated for similarity by comparisons in python. The quality of output texts was assessed by human radiologists using a score card.

## RSNA Template

The RSNA PICC insertion template can be found in [2]. Template items are listed in Table a. In the present study, three distinct cases were defined varying regarding eg anatomy, clinical information, and occurring complications. The impression had to be generated by the AI tool. "Patient ID", and "Study ID" were added as parameters for the present study, in addition to the template items contained in [2].

## ChatGPT Parameter Study

The defined cases were given as command file to ChatGPT [3] on 04 May 2023 and iterated (n = 5), producing 15 output cases in total. The command was set to:

"Write a radiology report which contains this exact information:".

No instruction was given on text structure, unlike before in [16]. The returned outputs were saved as txt-files. The previous study on distal radius fracture report drafting [16] relied on an earlier version of ChatGPT [18].

## Similarity Analysis in Python

An analysis of similarity between text output files was performed following a method used before relying on bag of words in python: cosine similarity between 0 and 1 of vectors given by key word occurrence in command files defining the indicator vector space [16], [19].

## Score Card Assessment

Table b contains the structure of the score card given to radiologists participating in this study as raters. In total, 5 questions had to be answered for each of the 15 output texts. For this, raters had to grade on an ordinal scale [+2, +1, 0, -1, -2] how much they agree / disagree with the following statements:

1. The report contains all relevant information.

2. I agree with the report's structure.

3. It is apparent that the text was written by an AI text drafting tool.

4. I would send this text unchanged as report to the referring physician.

5. In this case, the AI tool would have saved me time in my documentation / administration work.

Agreement regarding Questions 1, 2, 4, and 5 expresses a positive view on the ability of ChatGPT. As part of study's design, Question 3 was deliberately worded to require disagreement from the rater for expressing a positive view on the ability of ChatGPT. Raters were blinded to the results of the other raters. In total, 8 raters participated, 6 board


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

certified radiologists, 2 residents. The total work experience averaged 22.5 years (min 6, max 49) for the board certified radiologists, with an average of 14.2 years within IR (min 1, max 34). Both residents were in their second year of residency training with 0.5 years in IR.

### Interrater Agreement and Reliability

For analysing the agreement and reliability between raters, a set of variables was calculated from the score card results, Table d. Each variable took values in the interval [-1, 1]. The approach followed the methodology used before in [16]. Three agreement measures were calculated: exact agreement, one-apart agreement, and weighted agreement with weights for ordinal scales defined in [20]. Chance-corrected interrater reliability variables for the present study included: Gwet's AC1/AC2 (unweighted/ weighted), the Brennan-Prediger coefficient, Conger's kappa (generalization of Cohen's kappa for multiple raters), Fleiss' kappa, and Krippendorff's Alpha. These coefficients can be defined as 1 - (1 - Po)/(1 - Pe), where Po and Pe are measures of observed and chance agreement, respectively. The different variables differ in the definition of Po and Pe; detailed formulas are given in [20].

Imbalance in the occurrences of certain (pairs of) scores in the overall crosstabulation matrix makes traditionally used kappa variables as well as Krippendorff's Alpha prone to low reliability values. This paradox is further explained in [20]. Gwet's AC and the Brennan-Prediger coefficient are less influenced by this imbalance effect. Computations were made using the package provided in [21].

## Results

### Sample Output Text

Table c contains the output example for case 1, iteration 1, defined in Table a. It can be seen that in principle ChatGPT can draft a PICC line report following the required input from the command file. Throughout the present study, output text structure varied compared to the example of Table c. ChatGPT repeatedly changed the contained section headings. A variation of output text structure was not seen before in [16] where text structure had been an explicit part of the command file.

### Text Similarity throughout the Parameter Study

Fig. 1 lists the headings of sections produced by ChatGPT and extracted from the 15 output files. With the exception of "Patient ID" and "Study ID", no section heading appears in all 15 iterations. The average values included for Question 2 (I agree with the report's structure) demonstrate some substantial variation between the 15 cases. Performance was particularly rated as poor whenever no section on complications was included. Within the set of 5 iterations for each of the three cases, score of Question 2 drops / increases whenever the section on complications is omitted / included by ChatGPT. Fig. 2 provides a similarity comparison on a finer level and shows the cosine similarity calculated using bags of words. The comparison between the command files shows a 3 × 3 matrix with the main diagonal taking the max value of 1, comparing the command files with themselves. Pairwise similarity between different command files lies between 0.75 and 0.80. The comparison between command files and output files is plotted as a 3 × 15 grid. As before in [16], similarity exhibits plateaus of grid size 1 × 5 along the main diagonal, resulting from comparison between each of the command files to the 5 corresponding output files. Outside these three plateaus, similarity drops substantially. This pattern was seen before in the previous study. It demonstrates again that ChatGPT has the ability to adjust its output to minor changes in the command file. Remarkably, the similarity of the ChatGPT output from one case to a command file of a different case is on average not much lower than the similarity between the respective command files. Accordingly, not much similarity is lost when we move from a command file to the output. While the 1 × 5 similarity plateaus were highly homogeneous in the previous study, now on-plateau similarity varies markedly between values from 0.89 to 0.97. This, equally as before the change in text structure in Fig. 1, is new compared to [16] and can be attributed to the omission of prescribed text structure in the command file in the present study.

### Output Text Quality in Score Card Assessment

Fig. 3 plots the distribution of the rater responses per question. Fig. 4 shows the average rater response with one standard deviation as error bar. Table d contains in its first panel the mode, median, range, mean and standard deviation. Overall, raters agreed with the statements offered in Questions 1, 2, 4, and 5; while disagreeing regarding Question 3. This can be interpreted as a clear positive statement about the quality of the AI generated PICC insertion reports. Strong agreement was the most frequently given answer (strong disagreement in case of question 3 which was deliberately worded to require disagreement for a positive statement about the ChatGPT capabilities). By overall


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

rater opinion, all relevant information was included (Question 1) in an agreeable text structure (Question 2). Raters generally disagreed with the statement that the drafts were clearly identifiable as AI-generated. This response does not establish equivalence or indistinguishability relative to human-written reports, because a controlled human-report comparison was not performed. Question 4, whether the text draft could be sent out unchanged, saw a minor drop in mean agreement, as compared to the three previous questions. This indicates that raters would have considered editing the text draft manually before sending it. Raters generally agreed that the tool would have saved time in their documentation or administration work. This finding reflects anticipated benefit; actual documentation time was not measured. Essential points raised by raters in their comments concerned text structure and handling of medical complications. Note that complications already influenced results of Question 2 and Fig. 1. Raters missed medical treatment suggestions which should have been included by ChatGPT in the PICC insertion report for the referring doctor by their opinion.

## Rater Agreement and Interrater Reliability

An observation already made in [16] is confirmed by this study in Fig. 5 (standard deviation among raters, plotted over absolute rater mean; only negative means obtained under question 3 which required disagreement for a positive statement). Whenever texts are assessed to be of greater quality (abs(mean) approaches 2.0), variation between raters drops (standard deviation approaches 0.0). However, once quality is imperfect (abs (mean) approaches 0), there is an increasing variation between the raters' expression of lack of agreement (standard deviation approaches 1.8). The point scatter in Fig. 5 can be interpolated linearly by regression analysis, R2 = 0.830. On the view of the authors, the pattern in Fig. 5 reflects real life situations in eg case presentations where proportion of disagreement between radiologists may increase with greater need for discussion. Section 2 of Table d contains the calculated rater agreement. Question 4 which received the lowest absolute mean also shows the lowest agreement between raters for all three agreement variables. This reflects the pattern observed before in Fig. 5 and discussed above. By definition, the agreement variables increase in most cases for wider defined range when calculated per question: exact match < one-apart match < weighted match. Section 3 of Table d contains the calculated interrater reliability. Fair reliability was calculated for AC1 (unweighted / identity) and AC2 (weighted); as well as for weighted Conger's kappa, Fleiss' kappa, and Krippendorff's Alpha. The remaining measures led to only slight reliability. This range of values is more consistent than what was obtained before in [16] for the evaluation of distal radius fracture reports. Most remarkable is the drop of AC1/2 from (identity: substantial, weighted: almost Perfect) in [16] to (identity: fair, weighted: fair) in the present study. Brennan- Prediger also saw a drop compared to the levels obtained in [16]. The interrater reliability between individual raters is shown as pairwise heatmap in Fig. 6 for weighted AC2. Raters are sorted for decreasing AC2 when calculating it for k raters. It can be seen that pairwise reliability reaches values of up to 0.95. Weighted AC2 decreases to 0.87 for the first four raters. When the remaining four raters of the total 8 are added, AC2 decreases to 0.41. Independently of the ratings given, this finding too corresponds to real life experience according to which agreement between individual radiologists might well vary. Fig. 7 plots the reliability variables for each question. It demonstrates that AC2 and the Brennan-Prediger coefficient (except for Question 4) reached also in the present study greater values than the remaining variables, as before in [16]. The overall drop of AC2 and the Brennan- Prediger coefficient was effectively caused by question 4.

## Discussion

Due to the non-deterministic behaviour of ChatGPT, a parameter study was performed for each of the defined study cases (revisions n = 5). Unlike a previous study [16] in which text structure had been part of the input command, no required output text structure was given as part of the command file. As a result, the variation in text structure was stronger than in [16] (see Fig. 1). This drop in text similarity compared to [16] was also seen when calculating cosine similarity (see Fig. 2). Lack of reporting of complications as a separate report section by ChatGPT lowered scores on text structure (see Fig. 1). In the set of scores received from the raters, a clear pattern could be identified (linear regression, R2= 0.83) that standard deviation increases for lesser absolute mean, Fig. 5. This pattern reflects real life situations where proportion of disagreement between radiologists may increase with greater need for discussion. Pairwise analysis of interrater reliability in Fig. 6 showed that also as in real life agreement between individual raters varied (max AC2 0.95, min AC2 0.41).

Mathematics in medical diagnostics is a wide field [22] with potentially many options for optimising healthcare and hospital operations, not limited to automation of clinical documentation [23], [24]. AI tools might well find their way into application and support the interventional radiologist in his administration / documentation tasks. Measured time savings, if demonstrated in future studies, could be an important improvement [4], [5]. Patients, doctors, and hospital administrators would agree on that. Future work in this field will have to look deeper into ethical issues that may arise due to the application of ChatGPT in IR. One issue is whether professionals (radiologists, nurses etc.) trust AI-written reports. Also, patients may lose trust when they hear that reports are drafted using AI [25]. A second issue is how


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

responsibility is shared between humans and AI [26]: Should humans stay in the loop? And who takes responsibility if something goes wrong? Patient privacy also requires assessment when clinical information is processed by AI systems. Still, with continuing exposure of users and patients to AI tools and with steady improvements of technology and its ethical use, trust can be expected to grow.

## Limitations

The evaluation was limited to three predefined simulated cases, five outputs per case and eight raters. Repeated outputs from the same case do not represent independent clinical cases. The study assessed rater perceptions rather than measured workflow time, and it did not compare the drafts with a controlled set of human-written reports. Findings relate to the ChatGPT output generated on 4 May 2023 and should not be generalized to other model versions or routine clinical performance.

## Conclusion

ChatGPT generated PICC report drafts from the predefined case information. The participating radiologists generally rated the drafts positively and anticipated time savings, while noting editing needs and concerns about complications. These findings support further study of AI-assisted drafting but do not demonstrate measured time savings, equivalent quality or indistinguishability from human-written reports. Prospective workflow and clinical validation remain necessary.


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## Table a

Defined parameters for the simulated case study. RSNA template items [2] and common procedure text.

| RSNA item | Template content |
|---|---|
| Procedure | PICC insertion |
| Technique | Seldinger |
| Site | Right arm / left arm; basilic vein / brachial vein / cephalic vein |
| Catheter | Single-lumen / double-lumen / triple-lumen |
| PICC placement | Peripherally placed PICC line. The arm was prepped and draped in sterile fashion. Lidocaine 1% was used for local anesthetic. Under fluoroscopic and ultrasound guidance, the vein was patent and accessed with a micropuncture needle. A guide wire was then advanced into the vein. A vascular sheath was then advanced over a guide wire, and a PICC line was trimmed. The PICC line was then advanced into the central venous system. After confirmation of the catheter position, the catheter was sutured in place at the skin entry site. |
| Comparison | None |
| Findings | Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium. Complications: No immediate complication. |
| Fluoroscopy time | [min] |

## Table a continued

Case-specific input. Patient and Study IDs are additional simulated study parameters. The common PICC placement text above was included unchanged in each of the three case inputs.

| Parameter | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| Patient ID | KEHW7830 | OMSW2397247 | HBET29475 |
| Study ID | 2379430 | 395370 | 19482047 |
| Procedure | PICC insertion | PICC insertion | PICC insertion |
| Technique | Seldinger; US and fluoroscopy guidance | Seldinger; US and fluoroscopy guidance | Seldinger; venography and fluoroscopy guidance |
| Site | Right arm; brachial vein | Left arm; basilic vein | Left arm; cephalic vein |
| Catheter | Triple-lumen | Single-lumen | Double-lumen |
| Clinical information | 68 years, male, indication for chemotherapy | 82 years, female, coagulopathy preventing more permanent access | 75 years, female, total parenteral nutrition |
| Comparison | Chest x-ray from 22 Nov 2022 | Chest x-ray from 14 Sept 2021 | Chest x-ray from 24 Aug 2022 |
| Findings | Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium. | Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium. | Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium. |
| Complications | No immediate complication. | Arterial puncture with bleeding and haematoma | air embolus 2 ml |
| Fluoroscopy time | 2:36 min | 4:16 min | 3:23 min |
| Impression | AI auto-generated | AI auto-generated | AI auto-generated |


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## Table b

Radiology scoring card. Each case/iteration was rated on all five statements.

| Question | Statement |
|---|---|
| 1 | The report contains all relevant information. |
| 2 | I agree with the report's structure. |
| 3 | It is apparent that the text was written by an AI text drafting tool. |
| 4 | I would send this text unchanged as report to the referring physician. |
| 5 | In this case, the AI tool would have saved me time in my documentation / administration work. |

Case/iteration rows in the score card: n = 1; n = 2; n = … .

| Response | Score |
|---|---|
| Strongly agree | 2 |
| Agree | 1 |
| Undetermined | 0 |
| Disagree | -1 |
| Strongly disagree | -2 |

## Table c

ChatGPT output text for case 1, iteration 1 [3], using Table a and the RSNA PICC line template [2].

Radiology Report:

Patient ID: KEHW7830 Study ID: 2379430

Procedure: Peripherally inserted central catheter (PICC) insertion under ultrasound and fluoroscopy guidance

Indication: Chemotherapy

Technique: The right arm was prepped and draped in sterile fashion. Lidocaine 1% was used for local anesthesia. Under fluoroscopic and ultrasound guidance, the brachial vein was accessed with a micropuncture needle. A guide wire was advanced into the vein, followed by a vascular sheath. A triple-lumen PICC line was trimmed and advanced into the central venous system. The catheter position was confirmed, and the catheter was sutured in place at the skin entry site.

Findings: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium.

Complications: No immediate complications were noted.

Fluoroscopy time: 2 minutes and 36 seconds.

Impression: Peripherally placed triple-lumen PICC line successfully inserted in the right arm brachial vein under fluoroscopic and ultrasound guidance. The catheter tip is positioned at the junction of SVC and right atrium. No immediate complications were noted.

Signed, [Radiologist Name]


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## Table d

Simple statistics of score card results, rater agreement and interrater reliability.

### 1. Simple Statistics of Score Card Results

| Question | Case | Mode | Median | Range | Mean | SD |
|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 2 | 1 | 1.8 | 0.40 |
| 1 | 2 | 2 | 2 | 4 | 1.43 | 0.92 |
| 1 | 3 | 2 | 2 | 4 | 1.35 | 1.06 |
| 2 | 1 | 2 | 2 | 4 | 0.90 | 1.55 |
| 2 | 2 | 2 | 2 | 4 | 1.23 | 1.23 |
| 2 | 3 | 2 | 2 | 4 | 1.63 | 1.07 |
| 3 | 1 | -2 | -2 | 3 | -1.48 | 0.84 |
| 3 | 2 | -2 | -1 | 4 | -1.13 | 1.21 |
| 3 | 3 | -2 | -1 | 4 | -0.95 | 1.30 |
| 4 | 1 | 1, 2 | 1 | 4 | 0.40 | 1.58 |
| 4 | 2 | 2 | 1 | 4 | 0.53 | 1.57 |
| 4 | 3 | 2 | 2 | 4 | 0.60 | 1.66 |
| 5 | 1 | 2 | 2 | 4 | 1.45 | 1.09 |
| 5 | 2 | 2 | 2 | 4 | 1.15 | 1.19 |
| 5 | 3 | 2 | 2 | 4 | 0.98 | 1.33 |

### 2. Rater Agreement in Score Card Results

| Match | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| Exact | 0.49 | 0.49 | 0.35 | 0.25 | 0.39 |
| One-apart | 0.88 | 0.73 | 0.74 | 0.49 | 0.64 |
| Weighted | 0.87 | 0.78 | 0.81 | 0.62 | 0.77 |

### 3. Interrater Reliability over All Questions and Cases

| Coefficient | Value | Weights | Po | Pe | Confidence interval | Landis-Koch |
|---|---|---|---|---|---|---|
| AC1 | 0.27 | identity | 0.39 | 0.17 | 0.21 – 0.32 | Fair |
| AC2 | 0.41 | weighted | 0.77 | 0.61 | 0.30 – 0.52 | Fair |
| Brennan-Prediger | 0.19 | weighted | 0.77 | 0.72 | 0.08 – 0.30 | Slight |
| Brennan-Prediger | 0.24 | identity | 0.39 | 0.20 | 0.19 – 0.29 | Slight |
| Conger’s kappa | 0.33 | weighted | 0.77 | 0.66 | 0.23 – 0.44 | Fair |
| Conger’s kappa | 0.13 | identity | 0.39 | 0.30 | 0.08 – 0.17 | Slight |
| Fleiss’ kappa | 0.33 | weighted | 0.77 | 0.66 | 0.22 – 0.43 | Fair |
| Fleiss’ kappa | 0.11 | identity | 0.39 | 0.32 | 0.06 – 0.16 | Slight |
| Krippendorff’s alpha | 0.33 | weighted | 0.77 | 0.66 | 0.22 – 0.44 | Fair |
| Krippendorff’s alpha | 0.11 | identity | 0.39 | 0.32 | 0.06 – 0.16 | Slight |


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## Figures

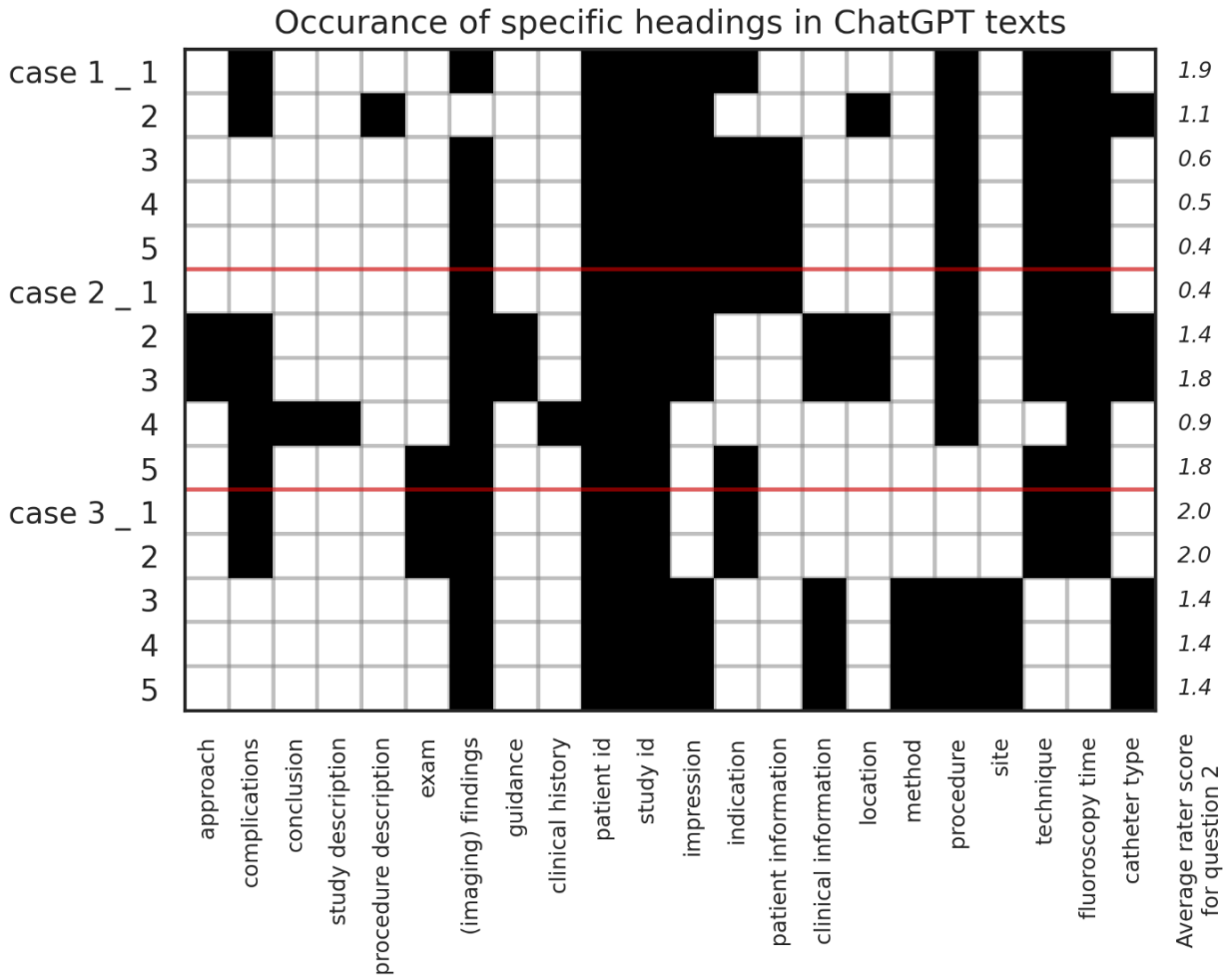



Figure 1. Section headings extracted from the 15 output files, sorted alphabetically by second word, with average rater score for Question 2: "I agree with the report's structure."

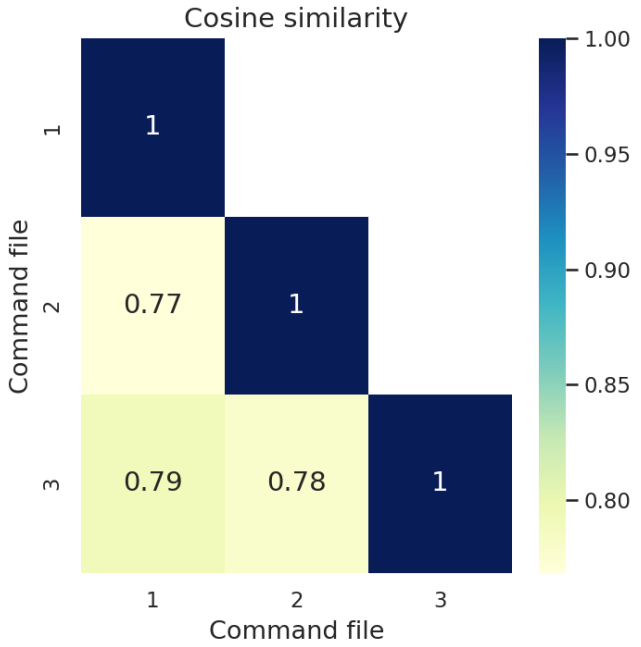



Figure 2A. Cosine similarity between the three command files.


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## Cosine Similarity and Rater Responses

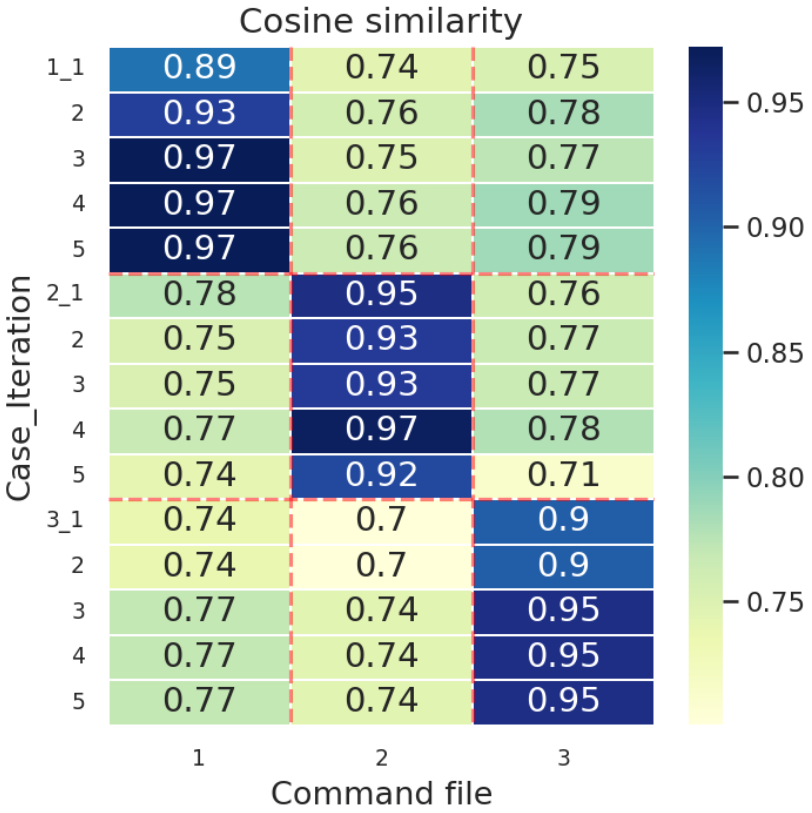



Figure 2B. Cosine similarity between command files and output files, computed using bag of words in Python.

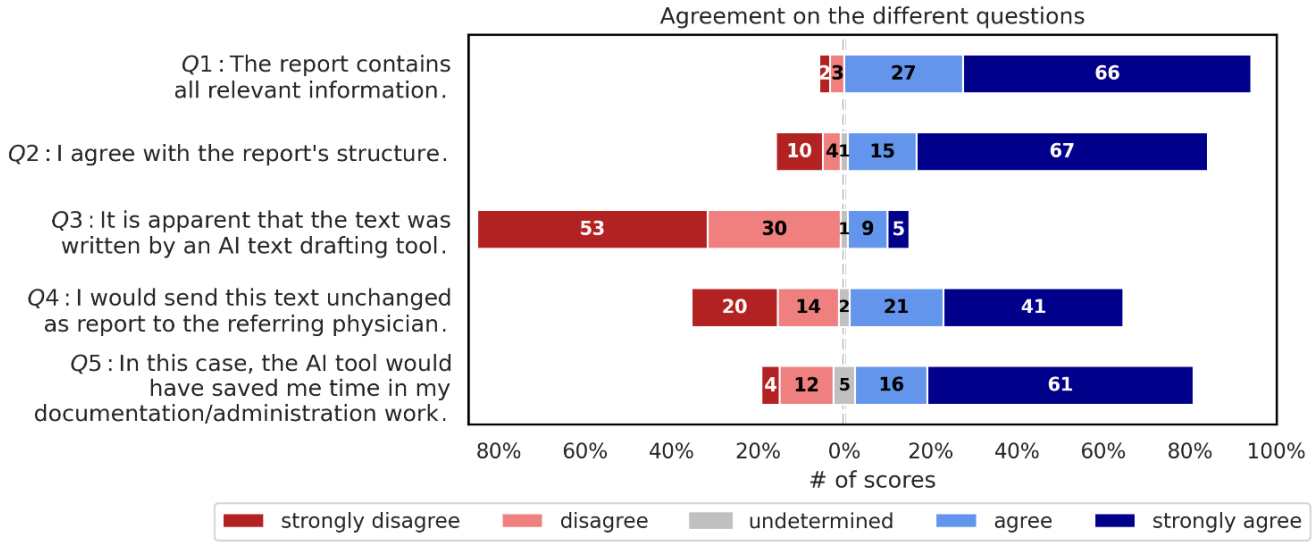



Figure 3. Distribution of agreement and disagreement by raters for each question.


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## Scores and Variation

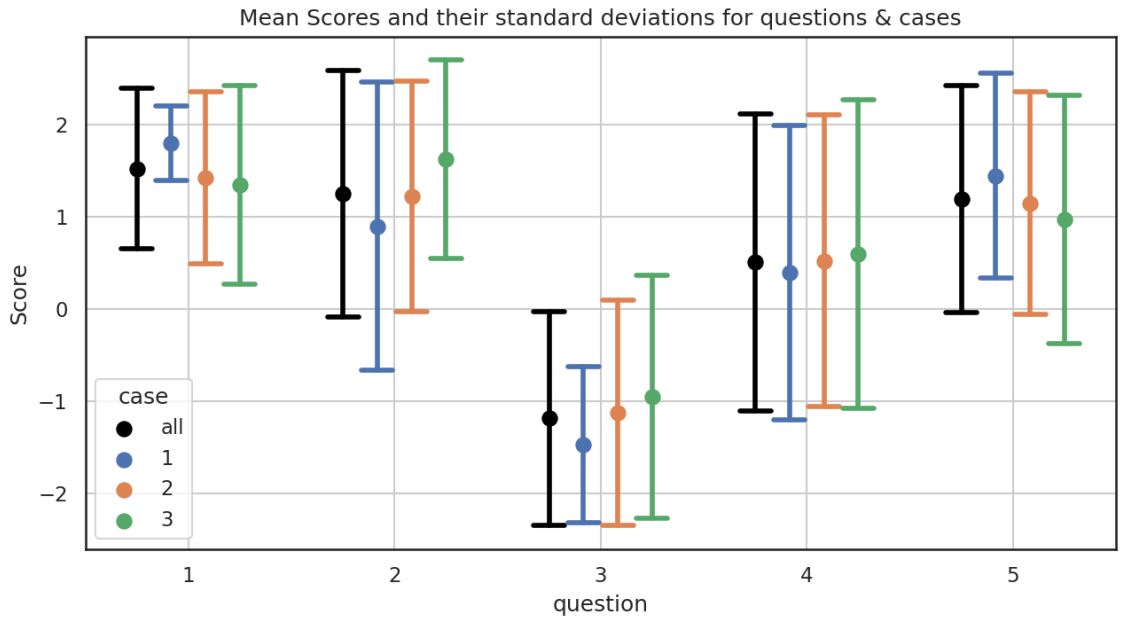



Figure 4. Mean score with error bars of ±1 standard deviation.

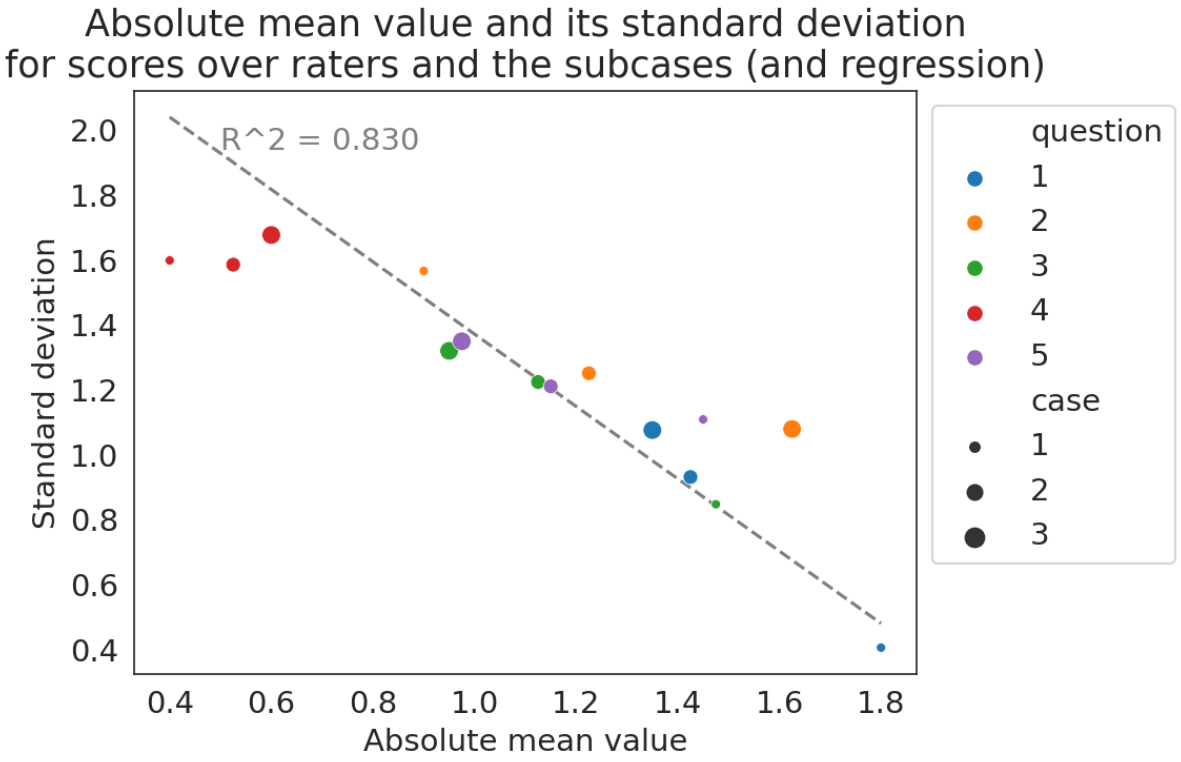



Figure 5. Standard deviation versus absolute mean score, aggregated per question and case (15 data points).


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## Interrater Reliability

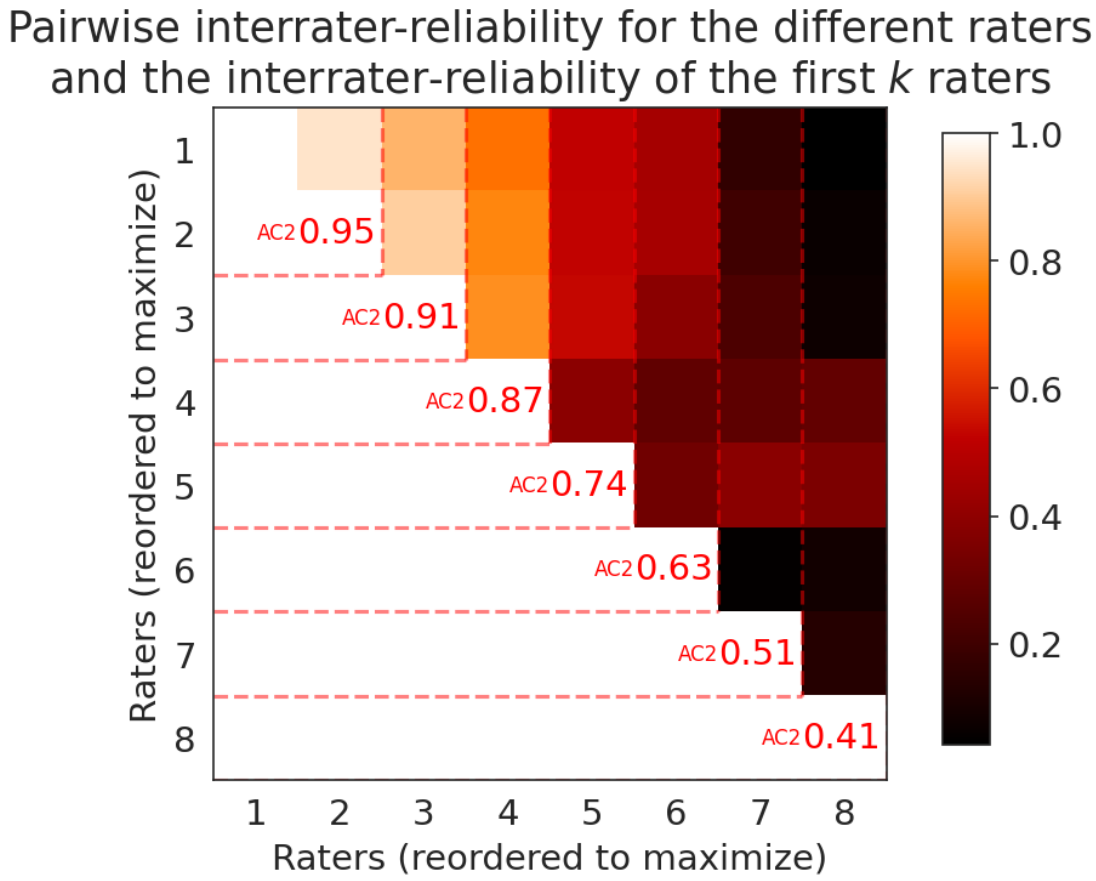



Figure 6. Pairwise interrater reliability heat map and reliability for the first k raters (red). Raters are sorted in descending order of Gwet's AC2 for successively larger groups.

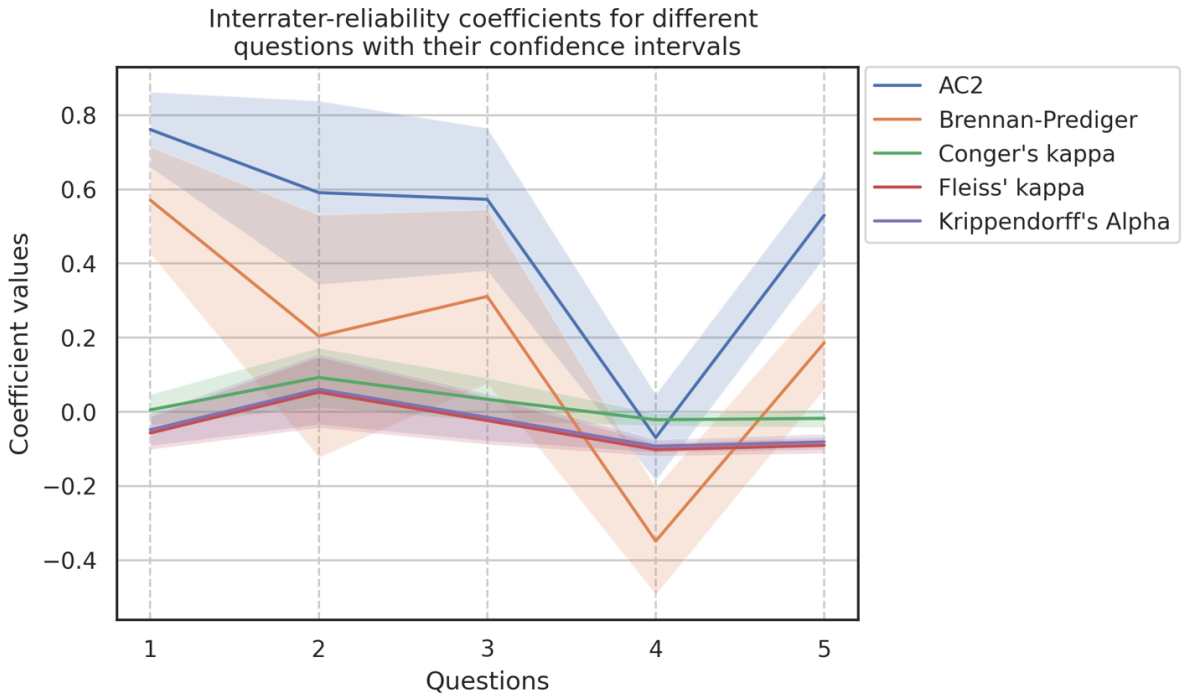



Figure 7. Weighted interrater reliability coefficients by question.


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## References


1. UVA Radiology and Medical Imaging, Ed., "What is interventional radiology?," Inside View (accessed 17 June 2023). https://blog.radiology.virginia.edu/interventional-radiologist-definition/
2. Medical College of Wisconsin, "PICC Insertion," RSNA RadReport, 2012 (accessed 3 May 2023). https://radreport.org/home/188/2012-05-29
3. OpenAI LLC, Ed., "ChatGPT — Release Notes (May 3)." (accessed 4 May 2023). https://help.openai.com/en/articles/6825453-chatgpt-release-notes
4. S. Woolhandler and D. U. Himmelstein, "Administrative work consumes one-sixth of u.s. physicians' working hours and lowers their career satisfaction," Int. J. Heal. Serv., vol. 44, no. 4, pp. 635–642, 2014. https://doi.org/10.2190/HS.44.4.a
5. S. M. Erickson, B. Rockwern, M. Koltov, R. M. McLean, and the Medical Practice and Quality Committee of the American College of Physicians. Putting Patients First by Reducing Administrative Tasks in Health Care: A Position Paper of the American College of Physicians. Ann Intern Med. 2017;166(9):659–661. https://doi.org/10.7326/M16-2697
6. M. Henderson, "Radiology Facing a Global Shortage," RSNA News, 2023. https://www.rsna.org/news/2022/may/global-radiologist-shortage
7. G. Sutherland, N. Russell, R. Gibbard, and A. Dobrescu, The Value of Radiology , Part II - The Conference Board of Canada, no. June. Ottawa, CAN, 2019.
8. K. Zuse, "Aus mechanischen Schaltgliedern aufgebautes Speicherwerk," DE924107, 1937.
9. A. M. Turing, "I.-Computing machinery and intelligence," Mind - A Q. Rev. Psychol. Philos., vol. 236, pp. 433– 460, 1950. https://doi.org/10.1093/mind/LIX.236.433
10. J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence," 1955 (accessed 30 October 2021). https://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf
11. D. Glowacka, A. Howes, J. P. Jokinen, A. Oulasvirta, and Ö. Şimşek, "RL4HCI: Reinforcement Learning for Humans, Computers, and Interaction," Ext. Abstr. 2021 CHI Conf. Hum. Factors Comput. Syst., pp. 1–3, 2021. https://doi.org/10.1145/3411763.3441323
12. R. Doshi, K. Amin, P. Khosla, S. Bajaj, S. Chheang, and H. P. Forman, "Utilizing Large Language Models to Simplify Radiology Reports : a comparative analysis," medRxiv Prepr., 2023. https://doi.org/10.1101/2023.06.04.23290786
13. R. Bhayana, S. Krishna, and R. R. Bleakney, "Performance of ChatGPT on a Radiology Board-style Examination : Insights into Current Strengths and Limitations," Radiology, vol. 307, no. 5, p. e230582, 2023. https://doi.org/10.1148/radiol.230582
14. Q. Lyu, J. Tan, M. E. Zapadka, J. Ponnatapura, C. Niu, K. J. Myers, G. Wang, and C. T. Whitlow, "Translating Radiology Reports into Plain Language using ChatGPT and GPT-4 with Prompt Learning: Promising Results, Limitations, and Potential," Vis. Comput. Ind. Biomed. Art, vol. 6, no. 9, pp. 1–10, 2023. https://doi.org/10.1186/s42492-023-00136-5
15. M. Barat, P. Soyer, and A. Dohan, "Appropriateness of Recommendations Provided by ChatGPT to Interventional Radiologists," Can. Assoc. Radiol. J., pp. 1–6, 2023. https://doi.org/10.1177/08465371231170133
16. W. A. Bosbach, J. F. Senge, B. Nemeth, S. H. Omar, M. Mitrakovic, C. Beisbart, A. Horvath, J. T. Heverhagen, and K. Daneshvar, "Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier," Curr. Probl. Diagn. Radiol., 2023. https://doi.org/10.1067/j.cpradiol.2023.04.001
17. W. A. Bosbach. Online supplement to manuscript: "Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier." Zenodo. 2023. https://doi.org/10.5281/zenodo.7908791
18. OpenAI LLC, Ed., "ChatGPT — Release Notes (Jan 9)." (accessed 11 January 2023). https://help.openai.com/en/articles/6825453-chatgpt-release-notes
19. E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, "More than bags of words: Sentiment analysis with word embeddings," Commun. Methods Meas., vol. 12, no. 2–3, pp. 140–157, 2018. https://doi.org/10.1080/19312458.2018.1455817
20. K. L. Gwet, Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Gaithersburg, MD (USA): Advanced Analytics, LLC, 2014.
21. K. Gwet and A. Fergadis, "irrCAC - Chance-corrected Agreement Coefficients," 2023 (accessed 5 March 2023). https://irrcac.readthedocs.io/en/latest/index.html
22. W. A. Bosbach, J. F. Senge, and P. Dlotko, Eds., "2022 Proceedings of the 4th International Conference on Trauma Surgery Technology: Mathematics in medical diagnostics," 2022, pp. 1–36. https://doi.org/10.5281/zenodo.7191419
23. W. A. Bosbach, M. Heinrich, R. Kolisch, and C. Heiss, "Maximization of Open Hospital Capacity under Shortage of SARS-CoV-2 Vaccines-An Open Access, Stochastic Simulation Tool," Vaccines, vol. 9, no. 6, p. 546, 2021. https://doi.org/10.3390/vaccines9060546
24. W. A. Bosbach, "Open-access supplement: Maximisation of open hospital capacity under shortage of SARS-CoV-2 vaccines," zenodo, 2021. https://doi.org/10.5281/zenodo.4589333
25. J. J. Hatherley, "Limits of trust in medical AI," J. Med. Ethics, vol. 46, no. 7, pp. 478–481, 2020. https://doi.org/10.1136/medethics-2019-105935
26. M. Verdicchio and A. Perin, "When Doctors and AI Interact: on Human Responsibility for Artificial Risks," Philos. Technol., vol. 35, no. 11, pp. 1–28, 2022. https://doi.org/10.1007/s13347-022-00506-6



Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12

## Author Information and Declarations

Correspondence: Wolfram A. Bosbach · WolframAndreas.Bosbach@Insel.CH

Jan F. Senge · https://orcid.org/0000-0002-2889-5489

Matthew T. McMurray · https://orcid.org/0009-0003-4710-8377

Fabian Haupt · https://orcid.org/0000-0002-2401-922X

Philipe S. Breiding · https://orcid.org/0000-0002-6985-0332

Claus Beisbart · https://orcid.org/0000-0003-2731-6200

Keivan Daneshvar · https://orcid.org/0000-0002-0345-0379

Gerd Nöldge · https://orcid.org/0009-0001-4186-4986

Frank Mosler · https://orcid.org/0000-0002-2039-4911

Wolfram A. Bosbach · https://orcid.org/0000-0003-4182-3516

Online supplement: study raw data · https://doi.org/10.5281/zenodo.8140755

Study preprint · https://doi.org/10.1101/2023.07.14.23292578

## Acknowledgements

The authors wish to thank for all the useful discussions leading to this manuscript.

## Funding

No funding was received for this work.

## Ethics Approval

Ethics approval was not required: no human or animal trials were performed, and no patient data were used.

## Conflict of Interest

The authors declare that there were no conflicts of interest within the meaning of the recommendations of the International Committee of Medical Journal Editors when the article was written.

## Disclaimer and Publisher's Note

The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Swiss J. Radiol. Nucl. Med. and/or the editor(s). Swiss J. Radiol. Nucl. Med. and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.


Open Access · CC BY 4.0 · © 2024 The Author(s) · Published by SJORANM GmbH, Switzerland
ISSN 2813-7221 · Swiss J. Radiol. Nucl. Med. 2024 7(2):1-14 · DOI 10.59667/sjoranm.v7i2.12