

1 ChatGPT may free time needed by the interventional radiologist for administration 2 / documentation: A study on the RSNA PICC line reporting template.

3
4 Jan F. Senge^{1,2}, Matthew T. McMurray³, Fabian Haupt³, Philippe S. Breiding⁴, Claus Beisbart^{5,6}, Keivan
5 Daneshvar³, Alois Komarek³, Gerd Nöldge³, Frank Mosler³, Wolfram A. Bosbach^{3*}
6

7 Affiliation:

8 [1] Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany

9 [2] Max-Planck Dioscuri Centre for Topological Data Analysis, Warsaw, Poland

10 [3] Department of Diagnostic, Interventional and Pediatric Radiology (DIPR), Inselspital, Bern University
11 Hospital, University of Bern, Switzerland

12 [4] University Institute of Diagnostic and Interventional Neuroradiology, Inselspital, Bern University
13 Hospital, University of Bern, Switzerland

14 [5] Institute of Philosophy, University of Bern, Bern, Switzerland

15 [6] Center for Artificial Intelligence in Medicine, University of Bern, Bern, Switzerland

16 *Correspondence: WolframAndreas.Bosbach@Insel.CH

17 Received: date; Accepted: date; Published: date
18
19

20 Abstract:

21 **Motive:** Documentation and administration, unpleasant necessities, take a substantial part of the working
22 time in the subspecialty of interventional radiology. With increasing future demand for clinical radiology
23 predicted, time savings from use of text drafting technologies could be a valuable contribution towards our
24 field.

25 **Method:** Three cases of peripherally inserted central catheter (PICC) line insertion were defined for the
26 present study. The current version of ChatGPT was tasked with drafting reports, following the Radiological
27 Society of North America (RSNA) template.

28 **Key results:** Score card evaluation by human radiologists indicates that time savings in documentation /
29 administration can be expected without loss of quality from using ChatGPT. Further, automatically generated
30 texts were not assessed to be clearly identifiable as AI-produced.

31 **Conclusions:** Patients, doctors, and hospital administrators would welcome a reduction of the time that
32 interventional radiologists need for documentation and administration these days. If AI-tools as tested in the
33 present study are brought into clinical application, questions about trust into those systems eg with regard
34 to medical complications will have to be addressed.
35
36

37 Introduction

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

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

51 Initial steps towards the computing technology required for AI in IR and elsewhere can be found in the work
52 of Konrad Zuse [8]. The development of AI using digital computers was first proposed in eg [9], [10]. As other
53 subfields of AI, natural language processing in combination with reinforcement learning has recently seen
54 some remarkable advances [3], [11]. This is a broader development, not limited to the tool applied in the
55 present study [12]. Regarding the application in radiology and elsewhere, the strengths of properly trained

56 language processing lie in the huge knowledge base that is made available [13], and in the ability to
57 communicate in different styles of language [14]. So far, rather few studies on AI based language processing
58 have been published in IR. One relevant study has demonstrated limitations concerning accuracy of
59 recommendations for IR procedures [15]. This result is similar to what we found in a previous study about
60 the handling of technical and medical information in report drafting for distal radius fracture [16], [17].
61 For evaluating the ability of ChatGPT to handle the RSNA PICC line template [2], we defined 3 distinct cases
62 and iterated those for a parameter study (n = 5). Output texts were evaluated for content similarity and
63 rated by 8 human radiologists. The main focus of the study was to determine if automation of text drafting
64 seems feasible and will save time of the interventional radiologist.

65 **Method and Materials**

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

70 **RSNA template**

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

75 **ChatGPT parameter study**

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

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

79
80
81
82 No instruction was given on text structure, unlike before in [16]. The returned outputs were saved as txt-
83 files. The previous study on distal radius fracture report drafting [16] relied on an earlier version of ChatGPT
84 [18].

85 **Similarity analysis in Python**

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

89 **Score card assessment**

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

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

98
99
100 Agreement regarding Questions 1, 2, 4, and 5 expresses a positive view on the ability of ChatGPT. As part of
101 study's design, Question 3 was deliberately worded to require disagreement from the rater for expressing a
102 positive view on the ability of ChatGPT. Raters were blinded to the results of the other raters.

103 In total, 8 raters participated, 6 board certified radiologists, 2 residents. The total work experience averaged
104 22.5 years (min 6, max 49) for the board certified radiologists, with an average of 14.2 years within IR (min 1,
105 max 34). Both residents were in their second year of residency training with 0.5 years in IR.

106 **Interrater agreement and reliability.**

107 For analysing the agreement and reliability between raters, a set of variables was calculated from the score
108 card results, Table d. Each variable took values in the interval [-1, 1]. The approach followed the
109 methodology used before in [16]. Three agreement measures were calculated: exact agreement, one-apart
110 agreement, and weighted agreement with weights for ordinal scales defined in [20]. Chance-corrected

111 interrater reliability variables for the present study included: Gwet's AC1/AC2 (unweighted/weighted), the
112 Brennan-Prediger coefficient, Conger's kappa (generalization of Cohen's kappa for multiple raters), Fleiss'
113 kappa, and Krippendorff's Alpha. These coefficients can be defined via $1 - \frac{1-P_o}{1-P_e}$, where of P_o and P_e are
114 measures of observed and chance agreement, respectively. The different variables only differ in the
115 definition of P_o and P_e , for detailed formulas see [20]. Imbalance in the occurrences of certain (pairs of)
116 scores in the overall crosstabulation matrix makes traditionally used kappa variables as well as
117 Krippendorff's Alpha prone to low reliability values. This paradoxon is further explained in [20]. Gwet's AC
118 and the Brennan-Prediger coefficient are less influenced by this imbalance effect. Computations were made
119 using the package provided in [21].
120

121 Results

122 Sample output text

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

128 Text similarity throughout the parameter study

129 Fig. 1 lists the headings of sections produced by ChatGPT and extracted from the 15 output files. With the
130 exception of "Patient ID" and "Study ID", no section heading appears in all 15 iterations. The average values
131 included for Question 2 (I agree with the report's structure) demonstrate some substantial variation
132 between the 15 cases. Performance was particularly rated as poor whenever no section on complications
133 was included. Within the set of 5 iterations for each of the three cases, score of Question 2 drops / increases
134 whenever the section on complications is omitted / included by ChatGPT.

135 Fig. 2 provides a similarity comparison on a finer level and shows the cosine similarity calculated using bags
136 of words. The comparison between the command files shows a [3, 3] matrix with the main diagonal taking
137 the max value of 1, comparing the command files with themselves. Pairwise similarity between different
138 command files lies between 0.75 and 0.80.

139 The comparison between command files and output files is plotted as a [3, 15] grid. As before in [16],
140 similarity exhibits plateaus of grid size [1, 5] along the main diagonal, resulting from comparison between
141 each of the command files to the 5 corresponding output files. Outside these three plateaus, similarity drops
142 substantially. This pattern was seen before in the previous study. It demonstrates again that ChatGPT has
143 the ability to adjust its output to minor changes in the command file. Remarkably, the similarity of the
144 ChatGPT output from one case to a command file of a different case is on average not much lower than the
145 similarity between the respective command files. Accordingly, not much similarity is lost when we move
146 from a command file to the output. While the [1, 5] similarity plateaus were highly homogeneous in the
147 previous study, now on-plateau similarity varies markedly between values from 0.89 to 0.97. This, equally as
148 before the change in text structure in Fig. 1, is new compared to [16] and can be attributed to the omission
149 of prescribed text structure in the command file in the present study.

150 Output text quality in scorecard assessment

151 Fig. 3 plots the distribution of the rater responses per question. Fig. 4 shows the average rater response with
152 one standard deviation as error bar. Table d contains in its first panel the mode, median, range, mean and
153 standard deviation.

154 Overall, raters agreed with the statements offered in Questions 1, 2, 4, and 5; while disagreeing regarding
155 Question 3. This can be interpreted as a clear positive statement about the quality of the AI generated PICC
156 insertion reports.

157 Strong agreement was the most frequently given answer (strong disagreement in case of question 3 which
158 was deliberately worded to require disagreement for a positive statement about the ChatGPT capabilities).
159 By overall rater opinion, all relevant information was included (Question 1) in an agreeable text structure
160 (Question 2). Raters overall disagreed with the statement that the output texts shown to them had been
161 apparently written by an AI tool; accordingly, raters would not identify them as written by AI rather than by
162 a human radiologist (Question 3). Question 4, whether the text draft could be sent out unchanged, saw a
163 minor drop in mean agreement, as compared to the three previous questions. This indicates that raters
164 would have considered editing the text draft manually before sending it. Question 5 received stronger

165 agreement again, which affirms that, under the view of the participating human radiologists, AI-based
166 automated text drafting will save time required in IR for administration / documentation.
167 Essential points raised by raters in their comments concerned text structure and handling of medical
168 complications. Note that complications already influenced results of Question 2 and Fig. 1. Raters missed
169 medical treatment suggestions which should have been included by ChatGPT in the PICC insertion report for
170 the referring doctor by their opinion.

171 **Rater agreement and interrater reliability**

172 An observation already made in [16] is confirmed by this study in Fig. 5 (standard deviation among raters,
173 plotted over absolute rater mean; only negative means obtained under question 3 which required
174 disagreement for a positive statement). Whenever texts are assessed to be of greater quality ($\text{abs}(\text{mean}) \rightarrow$
175 2.0), variation between raters drops (standard deviation \rightarrow 0.0). However, once quality is imperfect (abs
176 (mean) \rightarrow 0), there is an increasing variation between the raters' expression of lack of agreement (standard
177 deviation \rightarrow 1.8). The point scatter in Fig. 5 can be interpolated linearly by regression analysis, $R^2 = 0.830$.
178 On the view of the authors, the pattern in Fig. 5 reflects real life situations in eg case presentations where
179 proportion of disagreement between radiologists may increase with greater need for discussion.
180 Section 2 of Table d contains the calculated rater agreement. Question 4 which received the lowest absolute
181 mean also shows the lowest agreement between raters for all three agreement variables. This reflects the
182 pattern observed before in Fig. 5 and discussed above. By definition, the agreement variables increase in
183 most cases for wider defined range when calculated per question: exact match < one-apart match <
184 weighted match.

185 Section 3 of Table d contains the calculated interrater reliability. Fair reliability was calculated for AC1
186 (unweighted / identity) and AC2 (weighted); as well as for weighted Conger's kappa, Fleiss' kappa, and
187 Krippendorff's Alpha. The remaining measures led to only slight reliability. This range of values is more
188 consistent than what was obtained before in [16] for the evaluation of distal radius fracture reports. Most
189 remarkable is the drop of AC1/2 from (identity: substantial, weighted: almost Perfect) in [16] to (identity:
190 fair, weighted: fair) in the present study. Brennan-Prediger also saw a drop compared to the levels obtained
191 in [16].

192 The interrater reliability between individual raters is shown as pairwise heatmap in Fig. 6 for weighted AC2.
193 Raters are sorted for decreasing AC2 when calculating it for k raters. It can be seen that pairwise reliability
194 reaches values of up to 0.95. Weighted AC2 decreases to 0.87 for the first four raters. When the remaining
195 four raters of the total 8 are added, AC2 decreases to 0.41. Independently of the rates given, this finding too
196 corresponds to real life experience according to which agreement between individual radiologists might well
197 vary.

198 Fig. 7 plots the reliability variables for each question. It demonstrates that AC2 and the Brennan-Prediger
199 coefficient (except for Question 4) reached also in the present study greater values than the remaining
200 variables, as before in [16]. The overall drop of AC2 and the Brennan-Prediger coefficient was effectively
201 caused by question 4.

202

203 **Conclusions and future work**

204 In the present study, we tested ChatGPT [3] for its ability to draft IR reports after PICC line insertion. Reports
205 had to follow the current RSNA template [2] for three predefined study cases (Table a). Evaluation of the
206 report drafts by human radiologists led to an overall positive assessment. One main result is: time savings in
207 clinical administration / documentation of IR procedures can be expected from using ChatGPT (question 5).
208 Future work will have to assess further the expectable magnitude of time savings when compared to today's
209 form of report writing which does typically not use AI generated drafts.

210 Overall, raters did not identify the output texts as written by an AI tool; this indicates that reports written by
211 AI are for the raters indistinguishable from reports written by human radiologists (Question 3).

212 Due to the non-deterministic behaviour of ChatGPT, a parameter study was performed for each of the
213 defined study cases (revisions $n = 5$). Unlike our previous study [16] in which text structure had been part of
214 the input command, no required output text structure was given as part of the command file. As a result, the
215 variation in text structure was stronger than in [16] (see Fig. 1). This drop in text similarity compared to [16]
216 was also seen when calculating cosine similarity (see Fig. 2). Lack of reporting of complications as a separate
217 report section by ChatGPT lowered scores on text structure (see Fig. 1).

218 In the set of scores received from the raters, a clear pattern could be identified (linear regression, $R^2 = 0.83$)
219 that standard deviation increases for lesser absolute mean, Fig. 5. This pattern reflects real life situations

220 where proportion of disagreement between radiologists may increase with greater need for discussion.
221 Pairwise analysis of interrater reliability in Fig. 6 showed that also as in real life agreement between
222 individual raters varied (max AC2 0.95, min AC2 0.41).
223 Mathematics in medical diagnostics is a wide field [22] with potentially many options for optimising
224 healthcare and hospital operations, not limited to automation of clinical documentation [23], [24]. AI tools
225 might well find their way into application and support the interventional radiologist in his administration /
226 documentation tasks. Time savings, as can be expected from the results of the present study, would be an
227 important improvement [4], [5]. Patients, doctors, and hospital administrators would agree on that.
228 Future work in this field will have to look deeper into ethical issues that may arise due to the application of
229 ChatGPT in IR. One issue is whether professionals (radiologists, nurses etc.) trust AI-written reports. Also,
230 patients may lose trust when they hear that reports are drafted using AI [25]. A second issue is how
231 responsibility is shared between humans and AI [26]: Should humans stay in the loop? And who takes
232 responsibility if something goes wrong? Finally, the privacy of patients is an issue because reinforcement
233 learning uses input data to further train the model. Still, with continuing exposure of users and patients to AI
234 tools and with steady improvements of technology and its ethical use, trust can be expected to grow.
235
236

237 **Acknowledgements and funding:** The authors wish to thank for all the useful discussions leading to this
238 manuscript.

239 **Declaration of interests:** The authors declare no competing financial interests.

240 **Ethics approval:** not required

241 **Online supplement:** study raw data deposited under doi.org/10.5281/zenodo.8140755

242

243

244 Bibliography

- 245 [1] UVA Radiology and Medical Imaging, Ed., “What is interventional radiology?,” *Inside View*.
246 <https://blog.radiology.virginia.edu/interventional-radiologist-definition/> (accessed Jun. 17, 2023).
- 247 [2] Medical College of Wisconsin, “PICC Insertion,” *RSNA RadReport*, 2012.
248 <https://radreport.org/home/188/2012-05-29 00:00:00> (accessed May 03, 2023).
- 249 [3] OpenAI LLC, Ed., “ChatGPT — Release Notes (May 3).” [https://help.openai.com/en/articles/6825453-](https://help.openai.com/en/articles/6825453-chatgpt-release-notes)
250 [chatgpt-release-notes](https://help.openai.com/en/articles/6825453-chatgpt-release-notes) (accessed May 04, 2023).
- 251 [4] S. Woolhandler and D. U. Himmelstein, “Administrative work consumes one-sixth of u.s. physicians’
252 working hours and lowers their career satisfaction,” *Int. J. Heal. Serv.*, vol. 44, no. 4, pp. 635–642,
253 2014, doi: 10.2190/HS.44.4.a.
- 254 [5] S. M. Erickson, B. Rockwern, M. Koltov, R. M. Mclean, and M. Practice, “Putting Patients First by
255 Reducing Administrative Tasks in Health Care: A Position Paper of the American College of Physicians
256 Putting Patients First by Reducing Administrative Tasks in Health Care: A Position Paper of the
257 American College of Physicians,” *Ann. Intern. Med.*, vol. 166, no. 9, pp. 659–661, 2017, doi:
258 10.7326/M16-2697.
- 259 [6] M. Henderson, “Radiology Facing a Global Shortage,” *RSNA News*, 2023.
260 <https://www.rsna.org/news/2022/may/global-radiologist-shortage> (accessed May 16, 2023).
- 261 [7] G. Sutherland, N. Russell, R. Gibbard, and A. Dobrescu, *The Value of Radiology, Part II - The*
262 *Conference Board of Canada*, no. June. Ottawa, CAN, 2019.
- 263 [8] K. Zuse, “Aus mechanischen Schaltgliedern aufgebautes Speicherwerk,” DE924107, 1937
- 264 [9] A. M. Turing, “I.-Computing machinery and intelligence,” *Mind - A Q. Rev. Psychol. Philos.*, vol. 236,
265 pp. 433–460, 1950.
- 266 [10] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, “A Proposal For The Dartmouth Summer
267 Research Project On Artificial Intelligence,” 1955.
268 <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf> (accessed Oct. 30, 2021).
- 269 [11] D. Glowacka, A. Howes, J. P. Jokinen, A. Oulasvirta, and Ö. Azimsek, “RL4HCI: Reinforcement Learning
270 for Humans, Computers, and Interaction,” *Ext. Abstr. 2021 CHI Conf. Hum. Factors Comput. Syst.*, pp.
271 1–3, 2021, doi: 10.1145/3411763.3441323.
- 272 [12] R. Doshi, K. Amin, P. Khosla, S. Bajaj, S. Chheang, and H. P. Forman, “Utilizing Large Language Models
273 to Simplify Radiology Reports : a comparative analysis,” *medRxiv Prepr.*, 2023, doi:
274 10.1101/2023.06.04.23290786.

- 275 [13] R. Bhayana, F. S. Krishna, and R. R. Bleakney, "Performance of ChatGPT on a Radiology Board-style
276 Examination : Insights into Current Strengths and Limitations," *Radiology*, vol. 307, no. 5, p. e230582,
277 2023.
- 278 [14] Q. Lyu, J. Tan, M. E. Zapadka, J. Ponnatapura, C. Niu, K. J. Myers, G. Wang, and C. T. Whitlow,
279 "Translating Radiology Reports into Plain Language using ChatGPT and GPT-4 with Prompt Learning:
280 Promising Results, Limitations, and Potential," *Vis. Comput. Ind. Biomed. Art*, vol. 6, no. 9, pp. 1–10,
281 2023, doi: 10.1186/s42492-023-00136-5.
- 282 [15] M. Barat, P. Soyer, and A. Dohan, "Appropriateness of Recommendations Provided by ChatGPT to
283 Interventional Radiologists," *Can. Assoc. Radiol. J.*, pp. 1–6, 2023, doi: 10.1177/08465371231170133.
- 284 [16] W. A. Bosbach, J. F. Senge, B. Nemeth, S. H. Omar, M. Mitrovic, C. Beisbart, A. Horvath, J. T.
285 Heverhagen, and K. Daneshvar, "Ability of ChatGPT to generate competent radiology reports for
286 distal radius fracture by use of RSNA template items and integrated AO classifier," *Curr. Probl. Diagn.
287 Radiol.*, 2023, [Online]. Available: <https://doi.org/10.1067/j.cpradiol.2023.04.001>
- 288 [17] W. A. Bosbach, J. F. Senge, B. Nemeth, S. H. Omar, M. Mitrovic, C. Beisbart, A. Horvath, J. T.
289 Heverhagen, and K. Daneshvar, "Online supplement to manuscript: 'Ability of ChatGPT to generate
290 competent radiology reports for distal radius fracture by use of RSNA template items and integrated
291 AO classifier.' Current problems in diagnostic radiology (2023).," *zenodo*, 2023, doi:
292 10.5281/zenodo.7908791.
- 293 [18] OpenAI LLC, Ed., "ChatGPT — Release Notes (Jan 9)." [https://help.openai.com/en/articles/6825453-](https://help.openai.com/en/articles/6825453-chatgpt-release-notes)
294 [chatgpt-release-notes](https://help.openai.com/en/articles/6825453-chatgpt-release-notes) (accessed Jan. 11, 2023).
- 295 [19] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, "More than bags of
296 words: Sentiment analysis with word embeddings," *Commun. Methods Meas.*, vol. 12, no. 2–3, pp.
297 140–157, 2018.
- 298 [20] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of*
299 *agreement among raters*. Gaithersburg, MD (USA): Advanced Analytics, LLC, 2014.
- 300 [21] K. Gwet and A. Fergadis, "irrCAC - Chance-corrected Agreement Coefficients," 2023.
301 <https://irrcac.readthedocs.io/en/latest/index.html#> (accessed Mar. 05, 2023).
- 302 [22] W. A. Bosbach, J. F. Senge, and P. Dlotko, Eds., "2022 Proceedings of the 4th International Conference
303 on Trauma Surgery Technology: Mathematics in medical diagnostics," 2022, pp. 1–36. doi:
304 10.5281/zenodo.7191419.
- 305 [23] W. A. Bosbach, M. Heinrich, R. Kolisch, and C. Heiss, "Maximization of Open Hospital Capacity under
306 Shortage of SARS-CoV-2 Vaccines-An Open Access, Stochastic Simulation Tool," *Vaccines*, vol. 9, no. 6,
307 p. 546, 2021, doi: 10.3390/vaccines9060546.
- 308 [24] W. A. Bosbach, "Open-access supplement: Maximisation of open hospital capacity under shortage of
309 SARS-CoV-2 vaccines," *zenodo*, 2021, doi: 10.5281/zenodo.4589333.
- 310 [25] J. J. Hatherley, "Limits of trust in medical AI," *J. Med. Ethics*, vol. 46, no. 7, pp. 478–481, 2020, doi:
311 10.1136/medethics-2019-105935.
- 312 [26] M. Verdicchio and A. Perin, "When Doctors and AI Interact: on Human Responsibility for Artificial
313 Risks," *Philos. Technol.*, vol. 35, no. 11, pp. 1–28, 2022, doi: 10.1007/s13347-022-00506-6.
- 314
315

	RSNA template items [2]	Case 1	Case 2	Case 3
Patient (additional study parameter)		Patient ID KEHW7830 Study ID 2379430	Patient ID OMSW2397247 Study ID 395370	Patient ID HBET29475 Study ID 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 Left arm	Right arm	Left arm	Left arm
	Basilic vein Brachial vein Cephalic vein	Brachial vein	Basilic vein	Cephalic vein
Catheter	Single-lumen Double-lumen Triple-lumen	Triple-lumen	Single-lumen	Double -lumen
PICC placement	<p>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.</p>	<p>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.</p>	<p>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.</p>	<p>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.</p>
Clinical information		68 years, male, indication for chemotherapy	82 years, female, coagulopathy preventing more permanent access	75 years, female, total parenteral nutrition
Comparison	None	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.	Position: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium.
	Complications: No immediate complication.	Complications: No immediate complication.	Complications: Arterial puncture with bleeding and haematoma	Complications: air embolus 2 ml
Fluoroscopy time:	[min]	Fluoroscopy time: 2:36 min	Fluoroscopy time: 4:16 min	Fluoroscopy time: 3:23 min
Impression				

AI auto generated

318 Table b: radiology scoring card.

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

319

strongly agree	2
agree	1
undetermined	0
disagree	-1
strongly disagree	-2

320

321 Table c: output example case 1, iteration 1, generated by ChatGPT [3] for values defined in Table a for the RSNA PICC line template [2].
322

ChatGPT output text, case 1, iteration 1:
<p><i>Radiology Report:</i></p> <p><i>Patient ID: KEHW7830</i> <i>Study ID: 2379430</i></p> <p><i>Procedure: Peripherally inserted central catheter (PICC) insertion under ultrasound and fluoroscopy guidance</i></p> <p><i>Indication: Chemotherapy</i></p> <p><i>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.</i></p> <p><i>Findings: A final placement radiograph demonstrates the tip of the catheter at the junction of SVC and right atrium.</i></p> <p><i>Complications: No immediate complications were noted.</i></p> <p><i>Fluoroscopy time: 2 minutes and 36 seconds.</i></p> <p><i>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.</i></p> <p><i>Signed,</i> <i>[Radiologist Name]</i></p>

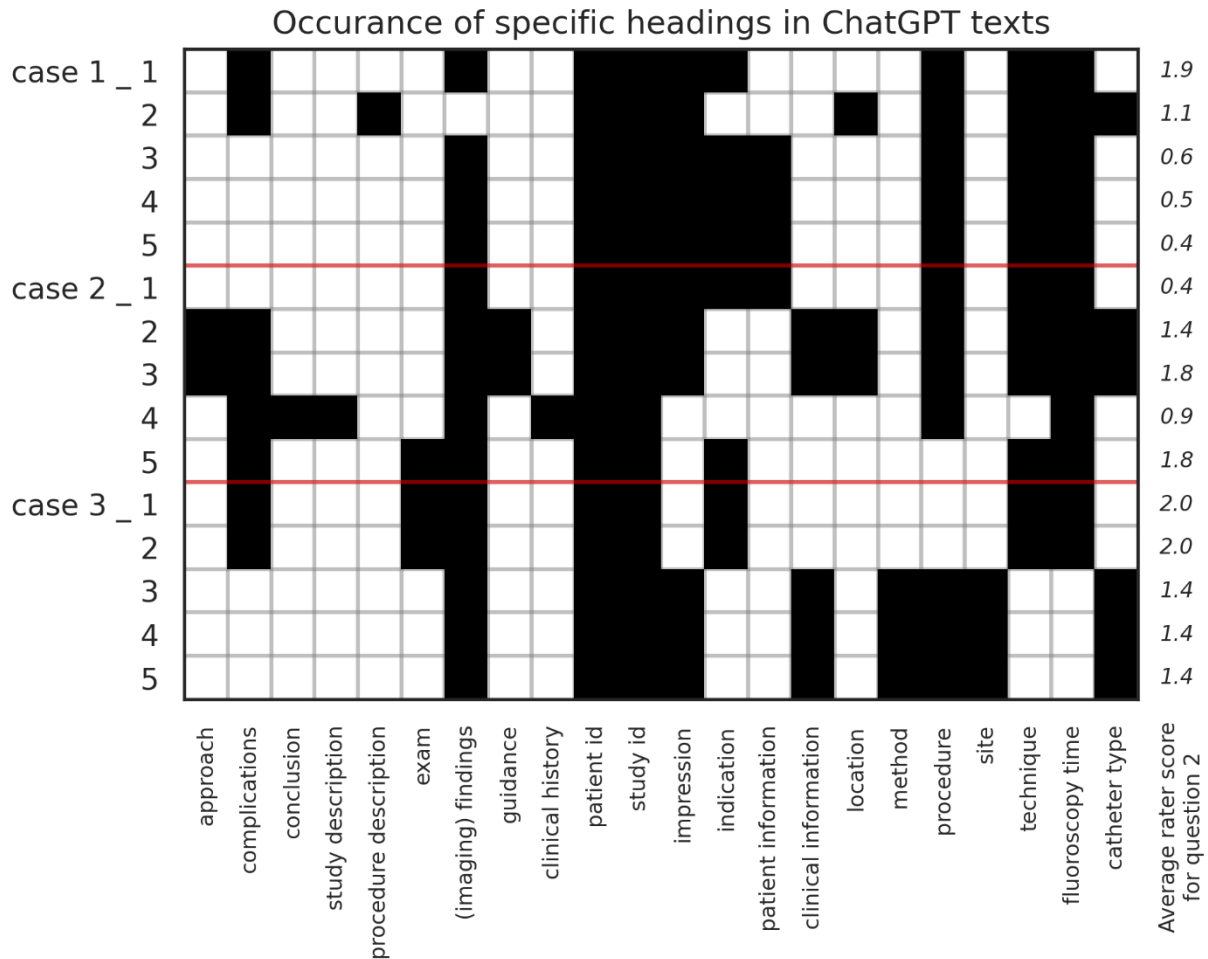
323

324 *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	stdev
1	1	2	2	1	1.8	0.40
	2	2	2	4	1.43	0.92
	3	2	2	4	1.35	1.06
2	1	2	2	4	0.90	1.55
	2	2	2	4	1.23	1.23
	3	2	2	4	1.63	1.07
3	1	-2	-2	3	-1.48	0.84
	2	-2	-1	4	-1.13	1.21
	3	-2	-1	4	-0.95	1.30
4	1	1, 2	1	4	0.40	1.58
	2	2	1	4	0.53	1.57
	3	2	2	4	0.60	1.66
5	1	2	2	4	1.45	1.09
	2	2	2	4	1.15	1.19
	3	2	2	4	0.98	1.33
2. rater agreement in score card results						
	match	Question 1	Question 2	Question 3	Question 4	Question 5
	exact match	0.49	0.49	0.35	0.25	0.39
	one-apart match	0.88	0.73	0.74	0.49	0.64
	weighted match	0.87	0.78	0.81	0.62	0.77
3. interrater reliability in score card results over all questions and cases						
Coefficient name	value	weights	P_o	P_e	confidence interval	Benchmark: Landis-Koch
AC1 (identity)AC2 (weighted)	0.27	identity	0.39	0.17	0.21 – 0.32	Fair
	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
	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
	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
	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
	0.11	identity	0.39	0.32	0.06 – 0.16	Slight

325

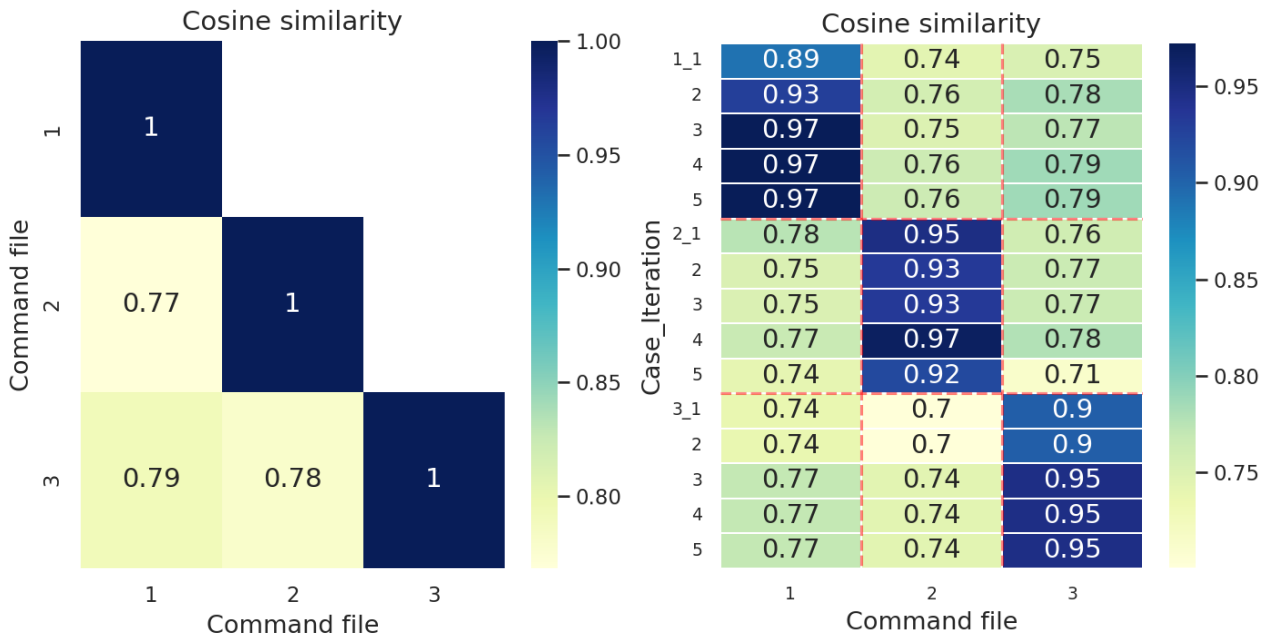
326



327

328 Fig. 1: section headings extracted from the 15 output files, sorted alphabetically by second word in heading, together with average
 329 value from raters for Question 2: I agree with the report's structure.

330

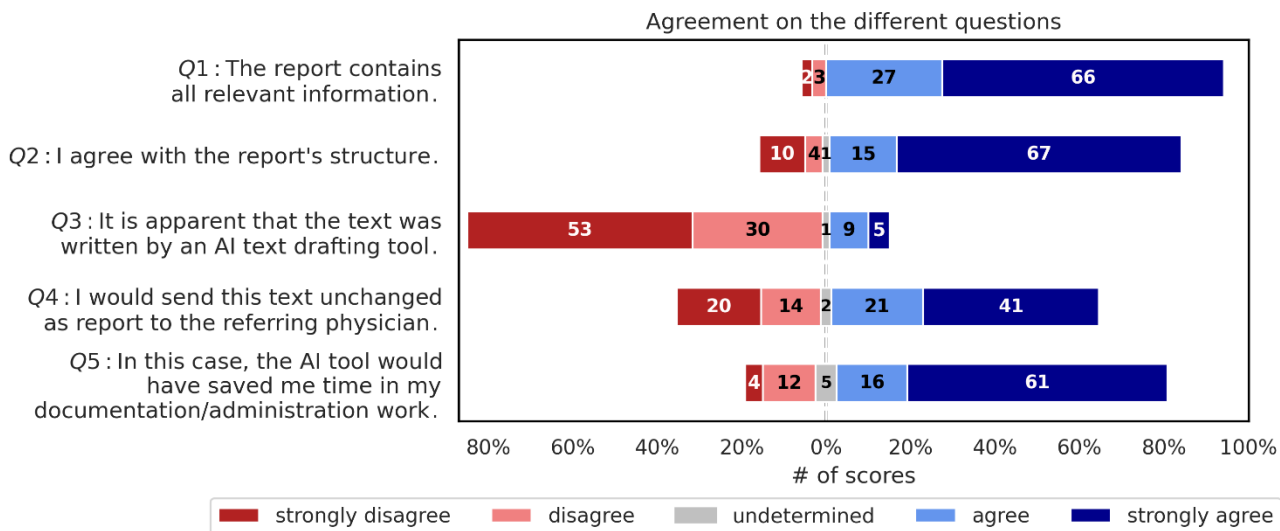


331

332 Fig. 2: cosine similarity matrix between command files, and between command files and output files, computed by bag of words in
 333 Python.

334

335

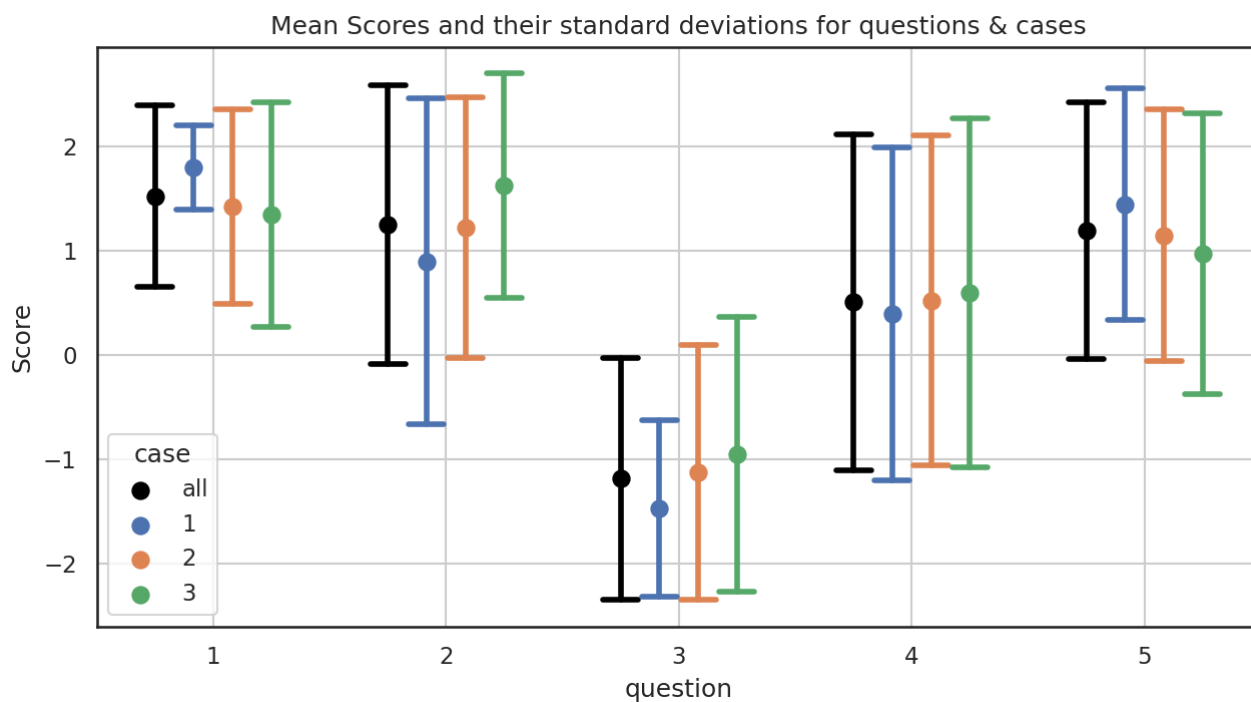


336

337 *Fig. 3: score card assessment with distribution of dis / agreement by raters per question.*

338

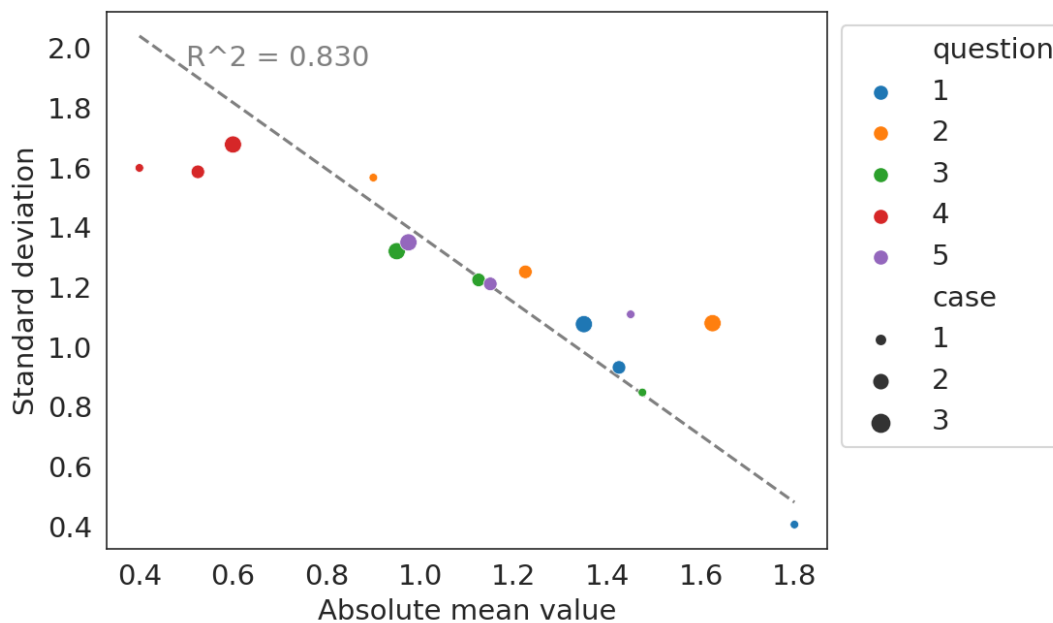
339



340

341 *Fig. 4: mean score with error bar of ± 1 standard deviation.*

342

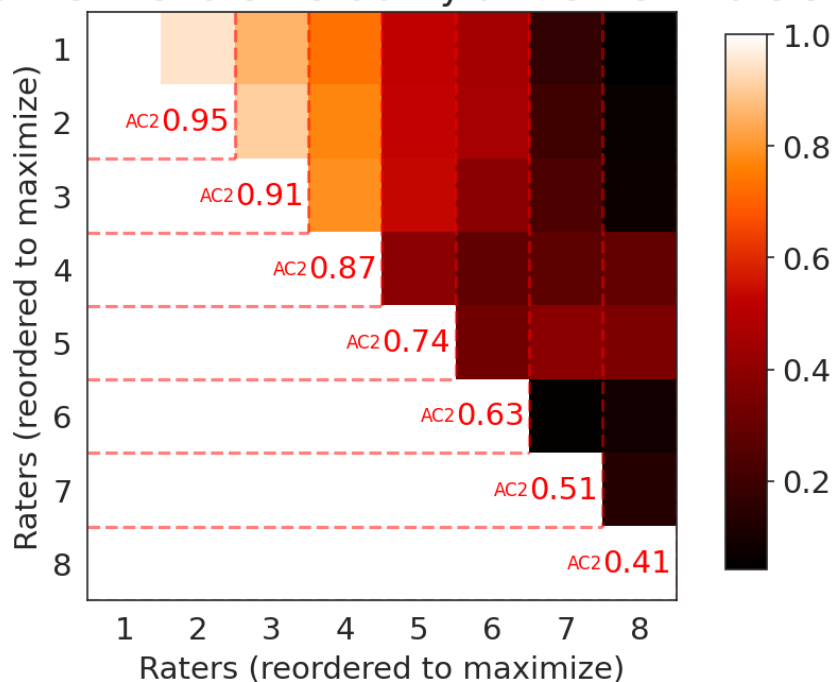


343

344 Fig. 5: standard deviation plotted over absolute mean, aggregated per question per case, 15 data points.

345

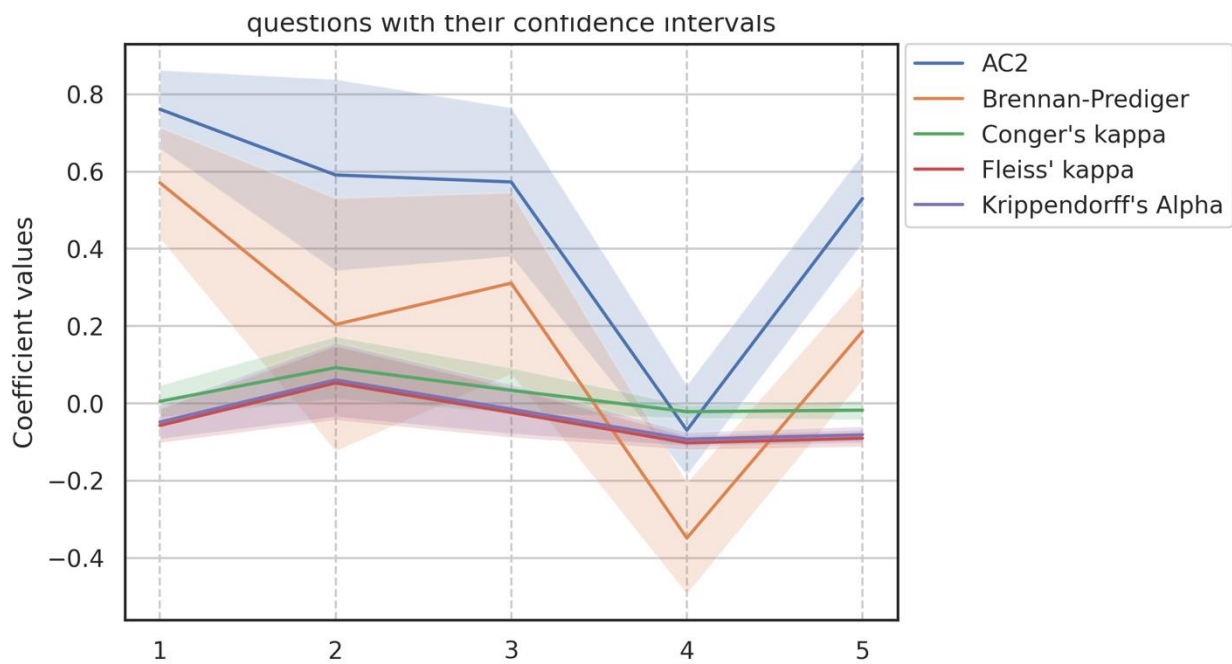
Pairwise interrater-reliability for the different raters and the interrater-reliability of the first k raters



346

347 Fig. 6: pairwise interrater reliability as heat map, as well as interrater reliability for group of the first k raters (red). The raters are
348 sorted for descending magnitude of Gwet's AC2 for greater group of raters.

349



350

351 *Fig. 7: weighted Interrater reliability variables per question.*

352

353