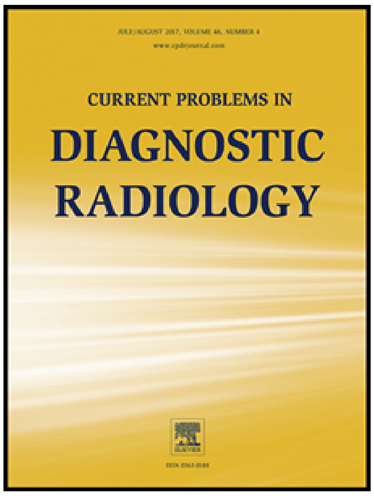
Journal Pre-proof



Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier

Wolfram A. Bosbach , Jan F. Senge , Bence Nemeth ,
Siti H. Omar , Milena Mitrakovic , Claus Beisbart ,
András Horváth , Johannes Heverhagen , Keivan Daneshvar

PII: S0363-0188(23)00052-X
DOI: https://doi.org/10.1067/j.cpradiol.2023.04.001
Reference: YMDR 1105

To appear in: *Current Problems in Diagnostic Radiology*

Please cite this article as: Wolfram A. Bosbach , Jan F. Senge , Bence Nemeth , Siti H. Omar , Milena Mitrakovic , Claus Beisbart , András Horváth , Johannes Heverhagen , Keivan Daneshvar , Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier, *Current Problems in Diagnostic Radiology* (2023), doi: https://doi.org/10.1067/j.cpradiol.2023.04.001

Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier

Wolfram A. Bosbach[1]*, Jan F. Senge[2,3], Bence Nemeth[1,4], Siti H. Omar[1,5], Milena Mitrakovic[1], Claus Beisbart[6,7], András Horváth[4], Johannes Heverhagen[1], Keivan Daneshvar[1]
Affiliation:
[1] Department of Diagnostic, Interventional, and Paediatric Radiology, University Hospital of Bern, Inselspital, University of Bern, Bern, Switzerland
[2] Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany
[3] Max-Planck Dioscuri Centre for Topological Data Analysis, Warsaw, Poland
[4] Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary
[5] Kuala Lumpur Hospital, Ministry of Health of Malaysia, Kuala Lumpur, Malaysia
[6] Institute of Philosophy, University of Bern, Bern, Switzerland
[7] Centre for Artificial Intelligence in Medicine, University of Bern, Bern, Switzerland
*Correspondence: WolframAndreas.Bosbach@Insel.CH
Received: date; Accepted: date; Published: date

Key points:

- The amount of acquired radiology imaging studies grows worldwide at a rapid pace.
- Automated text report drafting promises an increase of reporting quality and as well quantity.
- ChatGPT can be considered a substantial step forward towards the application of natural language processing in the drafting of radiological reports.
- Current limitations are the handling of technical / medical terminology.

Key words: natural language processing, report drafting, distal radius fracture.

Abstract:

Motive: The amount of acquired radiology imaging studies grows worldwide at a rapid pace. Novel information technology tools for radiologists promise an increase of reporting quality and as well quantity at the same time. Automated text report drafting is one branch of this development.
Method: We defined for the present study in total 9 cases of distal radius fracture. Command files structured according to a template of the Radiological Society of North America (RSNA) and to Arbeitsgemeinschaft Osteosynthese (AO) classifiers were given as input to the natural language processing tool ChatGPT. ChatGPT was tasked with drafting an appropriate radiology report.
Key results: A parameter study (n = 5 iterations) was performed. An overall high appraisal of ChatGPT radiology report quality was obtained in a score card based assessment. ChatGPT demonstrates the capability to adjust output files in response to minor changes in input command files. Existing shortcomings were found in technical terminology and medical interpretation of findings.
Conclusions: Text drafting tools might well support work of radiologists in the future. They would allow a radiologist to focus time on the observation of image details and patient pathology. ChatGPT can be considered a substantial step forward towards that aim.

Introduction

Clinical radiology in the Western world is facing a major challenge due to the demographic factors of an aging population, which is resulting in a rising demand for imaging services. Without a substantial increase in clinical imaging capacity, this growing demand will lead to longer-than-recommended waiting times for patients and will negatively impact patient outcomes [1]. Radiology as a service provider within interconnected modern medicine must not be overlooked when evaluating patient outcome [2]. The capacity increase will be achievable, so we believe, by greater facilities and by greater productivity of those. One tool that offers far reaching opportunities is information technology (IT) that is based upon artificial intelligence (AI) [3]–[5]. So far, AI research in radiology has mostly been directed at image analysis [6]–[8]. However, AI may also have the power to make a valuable contribution to other steps of the process, eg to

work list prioritisation [9] or to drafting reports for communicating results [10]. The processing of utterances in ordinary language and the drafting of texts have recently received much attention [11], [12]. Accordingly, there is a lively discussion on the extent to which AI can be used to draft eg scholarly work [13] or legal analyses [14]. It is also debated how AI may be integrated into education [15]. One important reason for AI technology's recent success in terms of user uptake is the ease in usability through simple user interfaces; an approach also known from eg data science projects [13].
In this paper, we use ChatGPT. This program was trained relying on Reinforcement Learning from Human Feedback [14]. The following three steps were taken [12]: First, an earlier language model, GPT-3.5 was fine-tuned using human conversations. Second, a reward model was built for reinforcement learning; for this, AI trainers ranked model outputs. Finally, a Proximal Policy Optimization Algorithm [15] was initialized with the fine-tuned model from step 1 and optimized using the reward model.
It appears firm at the moment that these novel tools will find their way into medical writing [15]. Quality of AI generated reports and their evaluation so far remain a challenge; parameters and metrics for the evaluation of report quality are still in their infancy, the main challenge being to capture the entirety of information presented by the imaging exam [16], [17]. In spite of these difficulties, language based AI tools possess great potential. Elaboration through speech is deeply rooted in medical communication between doctors. Text-based explanation of AI tool decision making in medical application can even be preferred by doctors over eg visualisation through salience maps [18]. AI-supported IT solutions eg in complex cancer reporting promise to improve quality and compliance of procedures, including eg recommendation for imaging findings [19]. The risk of losing information in the clinical process can be lowered by language competent IT which provides AI structured reporting [20].
The aim of this present study was to test and evaluate currently available AI text drafting tools in a typical radiological context. We chose the description of distal radius fracture as this study's focus pathology because this is a frequent and simple task. We refer the AI to the structure of a standard RSNA RadReport template [21] of the Radiological Society of North America (RSNA). Description of fragments and joint involvement follows the current Arbeitsgemeinschaft Osteosynthese (AO) classification [22]. Cases are defined by buzz words, see Table a. The output returned by the AI tool ChatGPT [12] is evaluated twofold. Python-based analyses are run for investigation of text similarity. Four radiologists specialising in MSK evaluate content by use of a score card, see Table b.

Method and Materials

In the present study, we test a language processing AI tool based on reinforcement learning and policy optimisation [14], [15] for its ability to draft competent radiology reports [12]. We define a total of 9 test cases of distal radius fracture, Table a. The input information for the report follows the structure of a typical RSNA template [21] in combination with the AO fracture classification [22].

RSNA template and AO classification

The RSNA template for (avulsion) fracture of the wrist / hand is published under [21], see step 1 in Table c. It is classically structured in exam information, findings, and impression. For the present study, a further style format parameter was added: merging of findings / impression vs keeping both sections separate.
Three sub / groups of AO distal radius fracture classification were chosen for the investigations in the present study. Three cases per sub / group were defined; giving a total of 9 cases for the present study (Table a). The study includes intra as well as extraarticular fractures, and gives different levels of eg displacement. The AO sub / group classifiers were [22]:

1. 2R3A1
    - Type: Radius, distal end segment, extraarticular fracture
    - Group: Radius, distal end segment, extraarticular, radial styloid avulsion fracture
2. 2R3A2.2
    - Type: Radius, distal end segment, extraarticular fracture
    - Group: Radius, distal end segment, extraarticular, simple fracture
    - Subgroup: dorsal displacement/tilt (Colles),
3. 2R3B1.1
    - Type: Radius, distal end segment, partial articular fracture
    - Group: Radius, distal end segment, partial articular, sagittal fracture
    - Subgroup: Involving scaphoid fossa

ChatGPT parameter study

The content of Table a was merged into one command file per case; the elements of the content were separated by commas with a generic command specification to write a radiology report in the beginning, example shown in step 2 of Table c. The command files were given as input to the text drafting tool [12]. Each command file was run for n = 5 iterations. Considering the additional study parameter on style (findings and impression separately vs impression only) and the total of 9 cases, the parameter study contains a total of 90 iterations. ChatGPT returned for each command file run a draft radiology report (step 3 in Table c).

Numerical text evaluation by Python

An assessment of drafted radiology reports with regard to possible omission of information or ability to reflect also minor changes from the input command files was performed. For that purpose, a bag of words tool was implemented in Python. Cosine similarity was extracted from the text files of the parameters study, range [0, 1]. Key word occurrence in command files was used as indicator vector space [23].

Radiological score card

For evaluation of text quality, the returned radiology report drafts were assessed independently by 4 expert radiologists. Three board certified radiologists with 17, 15, and 12 years of experience in radiology participated; as well as one radiology resident with 2 years of work experience.

Each of the 90 text drafts was scored with regard to 5 categories: correctness of exam information and fracture findings, suitability of impression, grammar, and style format. In each of the 5 categories, a score from a 5 point Likert scale [-2, -1, 0, 1, 2] was assigned; expressing the reviewing radiologist's level of dis / agreement, strong dis / agreement, or undetermination (Table b). During this process, the reviewers were blinded to the results of the other reviewers.

The obtained scores from the 4 radiologists were assessed for interrater reliability. Observable agreement of the ordinal scale is performed using three different agreement measures:

- exact agreement: identical scores from all reviewers
- one-apart agreement: at most two neighbouring scores from all reviewers
- weighted agreement: calculated by weights of Gwet [24] for ordinal scales, depending on agreement deviation.

Further, chance-corrected agreement measures are assessed:

- Gwet's AC1/AC2,
- Brennan-Prediger,
- Conger's kappa (generalization of Cohen's kappa for multiple raters),
- Fleiss' kappa,
- Krippendorff's alpha.

All chance corrected agreement coefficients are given by the same equation, see (Eq. 1):

$$1 - \frac{1 - P_o}{1 - P_e} \qquad \textit{(Eq. 1)}$$

Only the terms for observed agreement P_o and chance agreement P_e differ between different coefficients. These values are calculated from the scores of the different reviewers using eg [25].While the first three reliability coefficients are the most common ones, a problem arises in certain cases as mentioned in [26] due to high prevalence of a particular score. In [24], Gwet mentions that the problem of these so-called kappa paradoxes lies in the way agreement by chance P_e is defined. AC1 -in the weighted case AC2- as well as Brennen-Prediger tackle some of these effects and are more resistant of the effects of high singular score prevalence. For interpretation of the reliability coefficients, a cumulative interval membership probability (CIMP) approach is used with the Landis-Koch boundaries for assessing the results and 95% CIMP threshold [25].

Results

In the four sections of this results chapter, sample text results will be shown, followed by an analysis of similarity between the generated text files. In the third and fourth section, the quality evaluation by the radiology sore card will be shown, together with obtained interrater reliability.

Sample text results

Table c contains under step 3 a sample text result returned as radiology report draft for the command file of case 1 defined for an AO 2R3A1 fracture (Table a). The style obtained in this sample was maintained throughout the test runs of the present study. Dis / similarities between the returned drafts following changes to the command file are discussed in the next section. Table d and Fig. 4 will give greater details about reviewer reception.

Numerical evaluation of ChatGPT reports in Python

Fig. 1 shows the similarity matrix for the defined input command files (step 2 in Table c). The initial command line was omitted, so merged and separate findings / impression are both represented by the identical graph. Due to the in total 9 defined cases for the present study (Table a), a 9x9 matrix results. Only one half of the matrix is shown, considering the symmetry along the main diagonal. On the main diagonal, each command file is compared to itself; bag of words calculates a similarity of the maximum possible value 1. Three plateau fields of high similarity along the main diagonal are obtained, each [3, 3] large. Those three plateaus correspond to the similarity between the three cases defined each for the three AO sub / group classifiers (AO 2R3A1, AO 2R3A2.2, AO 2R3B1.1 in Table a). Within these three plateaus which are still in the blue region of the colour bar, similarity of no less than 0.79 is obtained. Beyond those, similarity drops further to values as low as 0.56 on the yellow region of the colour bar (bottom left corner).

For the n = 5 iterations run in the present study for each of the 9 command files, Fig. 2 and Fig. 3 show the [9, 45] similarity half matrix as obtained by bag of words (for style setting findings / impression separately, and merged respectively). The [45, 9] matrices again are symmetric along the main diagonal as it is the case above for the matrix in Fig. 1 when comparing command file similarity.

The total of 9 defined command files (3 each for AO 2R3A1, AO 2R3A2.2, AO 2R3B1.1 in Table a) also in Fig. 2 and Fig. 3 results in three plateaus of high similarity. In this case, the three plateaus of [15, 3] symmetric half matrices are located as before along the main diagonal.

The [5, 1] matrices making up the main diagonal all reached the maximum similarity value of 1. This holds true for Fig. 2, and Fig. 3. These fields stand for the comparison between the command file an the n = 5 iterations. The similarity value of 1.0 reflects the command part to write a radiology report which contains "*exact information*" (step 2 of Table c).

For findings / impression separately, the lowest similarity value within the [15, 3] plateaus is 0.79, equally 0.79 for findings / impression merged. The lowest off plateau value in both cases is 0.56 in the lower left corner. These values are identical to what was seen before in Fig. 1 for command files comparison, reflecting again the command to use exact information.

Summarising the observations from Fig. 1 to Fig. 3, bag of words in Python implementation demonstrates that text similarity reaches plateaus for the 3 command files each defined for the AO sub / groups AO 2R3A1, AO 2R3A2.2, AO 2R3B1.1 in Table a. This pattern is seen for the command files themselves, as well as the ChatGPT draft reports. A minor (or more pronounced) change of the command input file results in a small (or more pronounced) change in the obtained output files.

Radiology score card: overall assessed quality level

An analysis of the scores given by the 4 reviewers are shown for the 5 evaluation categories in Table d. The table lists for findings / impressions separately and merged the mode, median, range, mean, and standard deviation. The overall assessment of the quality of the report drafts was very positive. The lowest average score was obtained for the evaluation category "impressions suitable" with findings / impression separately at 1.08; still indicating overall reviewer response greater than simple agreement. The greatest average scores were obtained for "style format correctness" with 2.00 for both style types, findings / impression separate and merged. This reflects the reviewers' view on the ability of ChatGPT to generate radiology report drafts on a highly competent level. Step 3 in Table c shows in detail an example text from this present study.

When drafting findings / impression separately, the main criticism of reviewers concerned the category "suitability of impressions". ChatGPT returned impressions which reviewers deemed too extensive, step 3 in Table c. Instead a more concise impression section after the complete listing of findings would have been preferred. Under the scores for merged findings / impression, the lowest average score was obtained for the category "exam information correct". ChatGPT returned in this section of the parameter study repeatedly "*dorsoplantar (DV) and lateral views*" for the command line "*projection imaging DV and lateral".* This reduced the obtained score as seen as incorrect by reviewers. The value of the category "fracture findings correct" is with 1.68 less than 1.94 obtained under findings / impression separately. Reviewers scored lower partly because the report detail "*The acute onset of the injury and the rest of the bones and soft tissues are normal.*" was considered incorrect as well. An acute injury should, by their opinion, not reported as "*normal*".

Variation of reviewer scores when measured by standard deviation (Table d) increased for lower average scores. The lowest average scores ("impressions suitable" for findings / impression separately, and "exam information correct" for findings / impression merged) had the greatest standard deviation (1.08, and 1.20 respectively). The lowest standard deviation was obtained at 0.00 for the category of "style format correct" which received a reviewer score of 2.00 for both, findings / impression separately and merged.
Fig. 4 shows further the distribution bars of reviewer response. Each bar in Fig. 4 represents 180 scores: 9 defined cases x 5 iterations x 4 reviewers, for simplicity normalised to percent [%]. The most frequently picked answer is +2 (strong agreement) in each of the 5 categories for both, findings / impression separately and merged. The score of 0 (undetermined) was not assigned by reviewers in the present study. Correlating to the standard deviation of 0.00, exclusively +2 scores were given by reviewers in the category "style format correct".

Radiology score card: interrater reliability

Table d lists further the agreement between reviewers (exact, one-apart, and weighted) for the 5 evaluation categories. The resulting trends are consistent with the ones observable when analysing average and standard deviation. With dropping score averages, variation between reviewers increases, agreement measures decrease. The minimum agreement (by all three measures) is seen for "impressions suitable" when having findings / impression separate which had the lowest average score (1.08) in that style format. For findings / impression merged, the lowest agreement (according to one-apart and weighted match) is seen for "exam information correct" which received the lowest average score (1.39) in its respective style format. The best possible agreement of 1.00 is obtained for "style format correct".
The calculated interrater reliability measures from the scores of the 4 participating radiologists are shown in Table d. Considering the different interrater reliability coefficients for the unweighted (identity) as well as the weighted case with ordinal weights, the so-called kappa paradox is obtained in this study [26]. It results from the emphasis of the single score "strong agreement +2" (Fig. 4). Deviation in reviewer score choice reduces AC1+1 and Brennan-Prediger only marginally, while strongly reducint the kappa and alpha values. Low kappa and alpha (Conger's kappa, Fleiss' kappa, Krippendorff's Alpha) are matched with high agreement measured by AC1, AC2, and Brennan-Prediger. Using the Landis-Koch interpretation categories [25], the Brennan-Prediger as well as Gwet's AC1/AC2 coefficient show "Moderate" to "Substantial" for the unweighted as well as "Substantial" to "Almost Perfect" change-corrected for the weighted case. Kappa and alpha reach only "Slight" agreement.

Discussion

In the present manuscript, the natural language processing tool ChatGPT was tested for its ability to draft competent radiology reports. In total 9 input command files were defined with findings in distal radius fracture, following the structure of an RSNA template and AO classification (Table a). Quality assessment relied on a score card test in which 4 expert radiology reviewers participated (Table b). An overall high appraisal of ChatGPT radiology report quality was obtained (Table c). "Strong agreement" with the ChatGPT draft was the most frequently given score by human reviewers in this study. Criticism of reviewers focused on the length of the impression section; a more concise version would have been preferred instead. ChatGPT showed limitations in its ability to deal with technical/medical terminology. Dorsovolar (DV) given as input in command files was misinterpreted by ChatGPT as dorsoplantar. Another example was the putting of an "*acute onset of the injury*" into an overall "*normal*" context.
Text drafting tools might well support work of radiologists in the future. ChatGPT can be considered a substantial step forward towards that aim. Critical aspects in the future application of this technology will be eg potential for mass manipulation [27], but also substantial productivity increases [28].

Acknowledgements and funding: The authors wish to thank for all the useful discussions leading to this manuscript.
Declaration of interests: The authors declare no competing interests.
Ethics approval: not required

Online supplement:
ChatGPT parameter study with command files and (n = 5) iteration output files:
https://www.dropbox.com/sh/uubbgrkk88hbzpt/AADJyVpDS-0iNv7aYtvtC9qfa?dl=0

Bibliography

[1] 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.

[2] G. Schueller, "The role of the radiologist: When images save lives," *Imaging Med.*, vol. 2, no. 3, pp. 249–250, 2010, doi: 10.2217/iim.10.26.

[3] K. Zuse, "Aus mechanischen Schaltgliedern aufgebautes Speicherwerk," DE924107, 1937

[4] A. M. Turing, "I.-Computing machinery and intelligence," *Mind - A Q. Rev. Psychol. Philos.*, vol. 236, pp. 433–460, 1950.

[5] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence," 1955. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf (accessed Oct. 30, 2021).

[6] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nat. Rev. Cancer*, vol. 18, no. 8, pp. 500–510, 2018, doi: 10.1038/s41568-018-0016-5.

[7] H. Kaka and E. Zhang, "Artificial Intelligence and Deep Learning in Neuroradiology: Exploring the New Frontier," *Can. Assoc. Radiol. J.*, vol. 72, no. 1, pp. 35–44, 2021, doi: 10.1177/0846537120954293.

[8] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "DEEP CAPTIONING WITH MULTIMODAL RECURRENT NEURAL NETWORKS (M-RNN)," *arXiv*, vol. 1412, no. 6632v5, pp. 1–17, 2015, doi: 10.48550/arXiv.1412.6632.

[9] D. J. Winkel, T. Heye, T. J. Weikert, D. T. Boll, and B. Stieltjes, "Evaluation of an AI-Based Detection Software for Acute Findings in Abdominal Computed Tomography Scans: Toward an Automated Work List Prioritization of Routine CT Examinations," *Invest. Radiol.*, vol. 54, no. 1, pp. 55–59, 2019, doi: 10.1097/RLI.0000000000000509.

[10] M. H. Rezazade Mehrizi, P. van Ooijen, and M. Homan, "Applications of artificial intelligence (AI) in diagnostic radiology: a technography study," *Eur. Radiol.*, vol. 31, no. 4, pp. 1805–1811, 2021, doi: 10.1007/s00330-020-07230-9.

[11] S. Pichai, "An important next step on our AI journey." https://blog.google/technology/ai/bard-google-ai-search-updates/

[12] OpenAI LLC, Ed., "ChatGPT — Release Notes (Jan 9)." https://help.openai.com/en/articles/6825453-chatgpt-release-notes (accessed Jan. 11, 2023).

[13] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques,"* 4th ed. Hamilton, NZ, 2016.

[14] D. Glowacka, A. Howes, J. P. Jokinen, A. Oulasvirta, and Ö. Azimsek, "RL4HCI: Reinforcement Learning for Humans, Computers, and Interaction," *Ext. Abstr. 2021 CHI Conf. Hum. Factors Comput. Syst.*, pp. 1–3, 2021, doi: 10.1145/3411763.3441323.

[15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv*, vol. 1707, no. 06347, 2017, doi: 10.48550/arXiv.1707.06347.

[16] Z. Babar, T. van Laarhoven, F. M. Zanzotto, and E. Marchiori, "Evaluating diagnostic content of AI-generated radiology reports of chest X-rays," *Artif. Intell. Med.*, vol. 116, p. 102075, 2021, doi: 10.1016/j.artmed.2021.102075.

[17] F. Yu, M. Endo, R. Krishnan, I. Pan Md, A. Tsai, E. P. Reis, E. K. Ururahy, N. Fonseca, H. Min, H. Lee, Z. Shakeri, H. Abad, A. Y. Ng, C. P. Langlotz, V. Kumar Venugopal, and P. Rajpurkar, "Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation," *medRxiv*, p. 2022.08.30.22279318, 2022, [Online]. Available: https://www.medrxiv.org/content/10.1101/2022.08.30.22279318v1%0Ahttps://www.medrxiv.org/content/10.1101/2022.08.30.22279318v1.abstract

[18] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, "Producing radiologist-quality

reports for interpretable artificial intelligence," *arxiv.org*, pp. 1–7, 2018, doi: 10.48550/arXiv.1806.00340.
[19] B. C. Bizzo, R. R. Almeida, and T. K. Alkasab, "Computer-Assisted Reporting and Decision Support in Standardized Radiology Reporting for Cancer Imaging," *JCO Clin. Cancer Informatics*, no. 5, pp. 426–434, 2021, doi: 10.1200/cci.20.00129.
[20] J. T. Wu, A. Syed, H. Ahmad, A. Pillai, Y. Gur, A. Jadhav, D. Gruhl, L. Kato, M. Moradi, and T. Syeda-Mahmood, "AI Accelerated Human-in-the-loop Structuring of Radiology Reports," *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2020, pp. 1305–1314, 2020.
[21] H. Allen, M. D. Weintraub, B. G. Hansford, S. E. Stilwill, R. L. Leake, C. J. Hanrahan, B. Y. Chan, M. Soltanolkotabi, P. Kobes, and M. K. Mills, "XRay Wrist/Hand - Avulsion Fracture." https://radreport.org/home/50798/2019-10-28 15:08:23 (accessed Jan. 10, 2023).
[22] E. Meinberg, J. Agel, C. Roberts, M. Karam, and J. Kellam, "Fracture and Dislocation Compendium—2018-Orthopaedic Trauma Association, AO Foundation," *J. Orthop. Surg.*, vol. 31, no. Number 1 supplement, 2018.
[23] 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.
[24] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.
[25] K. Gwet and A. Fergadis, "irrCAC - Chance-corrected Agreement Coefficients," 2023. https://irrcac.readthedocs.io/en/latest/index.html# (accessed Mar. 05, 2023).
[26] A. R. Feinstein and D. V Cicchetti, "High agreement but low kappa: I. The problems of two paradoxes," *J. Clin. Epidemiol.*, vol. 43, no. 6, pp. 543–549, 1990.
[27] L. Illia, E. Colleoni, and S. Zyglidopoulos, "Ethical implications of text generation in the age of artificial intelligence," *Bus. Ethics, Environ. Responsib.*, no. July 2022, pp. 201–210, 2022, doi: 10.1111/beer.12479.
[28] E. Ernst, R. Merola, and D. Samaan, "Economics of Artificial Intelligence: Implications for the Future of Work," *IZA J. Labor Policy*, vol. 9, no. 4, 2019, doi: 10.2478/izajolp-2019-0004.

Table a: cases for defined parameter study

		AO 2R3A1, case 1	AO 2R3A1, case 2	AO 2R3A1, case 3	AO 2R3A2.2, case 1	AO 2R3A2.2, case 2	AO 2R3A2.2, case 3	AO 2R3B1.1, case 1	AO 2R3B1.1, case 2	AO 2R3B1.1, case 3
Exam	Projections	projection imaging DV and lateral,	projection imaging DV and lateral,	projection imaging DV and lateral,	projection imaging DV and lateral,	projection imaging DV and lateral,	projection imaging DV and lateral,	projection imaging DV and lateral,	projection imaging DV and lateral,	projection imaging DV and lateral,
	Joints	distal radius fracture,	distal radius fracture	distal radius fracture,	transvers distal radius fracture,	transvers distal radius fracture,	transvers distal radius fracture,	distal radius fracture,	distal radius fracture,	distal radius fracture,
	Side	Left,	Right,	Right,	Left,	Left,	Right,	Right,	Left,	Right,
Findings	Fracture description Fracture orientation Intra-articular extension Articular surface involvement	extraarticular, AO classification 2R3A1, fragment size 13 x 4 mm,	extraarticular, AO classification 2R3A1, fragment size 9 x 3 mm	extraarticular, AO classification 2R3A1, fragment size 15 x 5 mm	extraarticular, AO classification 2R3A2.2,	extraarticular, AO classification 2R3A2.2,	extraarticular, AO classification 2R3A2.2,	intraarticular, wedge at scaphoid fossa, fragment size 35 x 25 mm, AO classification 2R3B1.1,	intraarticular, wedge at scaphoid fossa, fragment size 39 x 17 mm, AO classification 2R3B1.1,	intraarticular, wedge at scaphoid fossa, fragment size 32 x 13 mm, AO classification 2R3B1.1,
	Fracture comminution	no comminution,	no comminution,	no comminution,	no comminution,	no comminution,	no comminution,	no comminution,	no comminution,	no comminution,
	Fracture fragment rotation	minimal volar rotation in DV,	minimal radial rotation in DV,	no rotation,	fragment dorsal tilt 30°,	fragment dorsal tilt 20°,	fragment dorsal tilt 50°,	No rotation,	No rotation,	fragment dorsal tilt 30°,
	Fracture fragment distraction	distraction radial direction 3 mm, distraction distal direction 7 mm,	distraction radial direction 4 mm, distraction distal direction 11 mm,	distraction radial direction 5 mm, no distraction distal,	distraction dorsal direction 6 mm, no distraction distal,	distraction dorsal direction 9 mm, distraction proximal direction 4 mm,	distraction dorsal direction 7 mm, no distraction distal,	intraarticular step of 5 mm,	intraarticular step of 8 mm,	intraarticular step of 4 mm,
	Soft tissue swelling	soft tissue swelling around fragment,	soft tissue swelling volar and palmar,	soft tissue swelling at radial styloid,	soft tissue swelling dorsal,	massive soft tissue swelling dorsal,	circular soft tissue swelling around the fracture site,	soft tissue swelling dorsal,	soft tissue swelling around the fracture,	soft tissue swelling dorsal,
	Fracture acuity	Subacute,	Acute,	fracture age indeterminate,	Acute,	Acute,	Subacute,	Acute,	Subacute,	Acute,
Additional information		rest normal	rest normal	rest normal	rest normal	rest normal	rest normal	rest normal	rest normal	rest normal
Impression	AI auto generated									

Table b: radiology scoring card

case	iteration	Exam information correct	Fracture findings correct	Impression suitable	Grammar correct	Style format correct
(see Table a)	Iteration: n = 1					
	n = 2					
	n = ..					

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

Table c: step 1 structure of RSNA template for avulsion fracture of the hand [21] *with additional style parameter for the present study with regard to merging of findings and impression, step 2 command file with sample values of case 1 of AO 2R3A1 from Table a, step 3 text results returned, showing iteration n=1 for both findings / impression separately and merged*

Step 1: Input command file structure RSNA template for avulsion fracture of the hand [21]	
Exam	• Projections • Joints • side
Findings	• Fracture description: [text] • Fracture comminution (Marked / Minimal / None) • Fracture orientation (Transverse / Oblique / Longitudinal) • Fracture fragment rotation (Marked / Minimal / None) • Intra-articular extension: [text] • Articular surface involvement (marked / minimal / none) • Fracture fragment distraction: [mm] • Soft tissue swelling (present / absent) • Fracture acuity (acute / subacute / chronic / age indeterminate) • Additional information
Impression	• [text]
Additional study parameter **Style format**	• Findings and impression, separately • Impression, only
Step 2: command file specification AO 2R3A1, case 1	
- findings / impression separately	*Write a radiology report structured into exam, findings and impression which contains this exact information, don't add notes at the bottom:*
- findings / impression merged	*Write a radiology report structured into exam and impression which contains this exact information, don't add notes at the bottom:*
Specified findings Example: AO 2R3A1, case 1 Following the structure of the RSNA template for distal radius	*projection imaging DV and lateral,* *distal radius fracture,* *Left,* *extraarticular,* *AO classification 2R3A1,* *fragment size 13 x 4 mm,* *no comminution,*

fracture [12]	*minimal radial rotation in DV, distraction radial direction 3 mm, distraction distal direction 7 mm, soft tissue swelling around fragment, Subacute, rest normal*

Step 3: sample text results
AO 2R3A1, case 1, iteration n=1

- findings / impression separately	*Exam: Projection imaging (DV and lateral) of the left distal radius* *Findings: Fracture of the distal radius, extraarticular. AO classification 2R3A1. Fragment size of 13 x 4 mm. No comminution. Minimal radial rotation in DV. Distraction in the radial direction of 3 mm, and in the distal direction of 7 mm. Soft tissue swelling around the fragment. Subacute injury.* *Impression: Distal radius fracture, extraarticular, AO classification 2R3A1. Subacute injury. No comminution and minimal radial rotation. Distraction in the radial direction of 3 mm and distal direction of 7 mm with soft tissue swelling around the fragment. Rest of the exam is normal.*
- findings / impression merged	*Exam: Projection imaging of the left distal radius was performed using both DV and lateral views.* *Impression: Subacute left extra-articular distal radius fracture, AO classification 2R3A1, with a fragment measuring 13 x 4 mm. No comminution is present. The fracture fragment has minimal radial rotation in the DV view and distraction of 3 mm in radial direction and 7 mm in distal direction. Soft tissue swelling is seen around the fragment, but the rest of the exam is normal.*

Table d: Central tendencies and variance of the 5-point Likert data of the responses, interpreted as ordinal data as well as interval data and their respective summary statistics.

1. simple statistics of score card results

style	category	mode	median	range	mean	std
findings / impressions separate	Exam information correct	2	2	3	1.96	0.33
	Fracture findings correct	2	2	1	1.94	0.24
	Impressions suitable	2	1	3	1.08	1.08
	Grammar correct	2	2	3	1.80	0.63
	Style format correct	2	2	0	2.00	0.00
only impression	Exam information correct	2	2	4	1.39	1.20
	Fracture findings correct	2	2	1	1.84	0.36
	Impressions suitable	2	2	3	1.68	0.62
	Grammar correct	2	2	1	1.68	0.47
	Style format correct	2	2	0	2.00	0.00

2. reviewer agreement in score card results

style	match	Exam information correct	Fracture findings correct	Impressions suitable	Grammar correct	Style format correct
findings / impressions separate	exact match	0.96	0.88	0.22	0.76	1.00
	one-apart match	0.98	1.00	0.63	0.92	1.00

	weighted match	0.98	0.99	0.78	0.94	1.00
only impression	exact match	0.60	0.69	0.61	0.51	1.00
	one-apart match	0.74	1.00	0.94	1.00	1.00
	weighted match	0.85	0.97	0.94	0.95	1.00

3. interrater reliability in score card results

Coefficient name	**value**	**weights**	P_o	P_e	**confidence interval**	**Benchmark: Landis-Koch**
AC1	0.69	identity	0.72	0.10	(0.65506, 0.72718)	Substantial
AC2	0.91	ordinal	0.94	0.27	(0.89467, 0.92811)	Almost Perfect
Brennan-Prediger	0.63	identity	0.72	0.25	(0.59021, 0.66905)	Moderate
	0.81	ordinal	0.94	0.66	(0.77631, 0.83872)	Substantial
Conger's kappa	0.12	identity	0.72	0.69	(0.08887, 0.14728)	Slight
	0.12	ordinal	0.94	0.93	(0.09986, 0.14433)	Slight
Fleiss' kappa	0.08	identity	0.72	0.70	(0.04532, 0.11561)	Slight
	0.07	ordinal	0.94	0.93	(0.04007, 0.0965)	Slight
Krippendorff's Alpha	0.08	identity	0.72	0.70	(0.04583, 0.11612)	Slight
	0.07	ordinal	0.94	0.93	(0.04059, 0.09702)	Slight

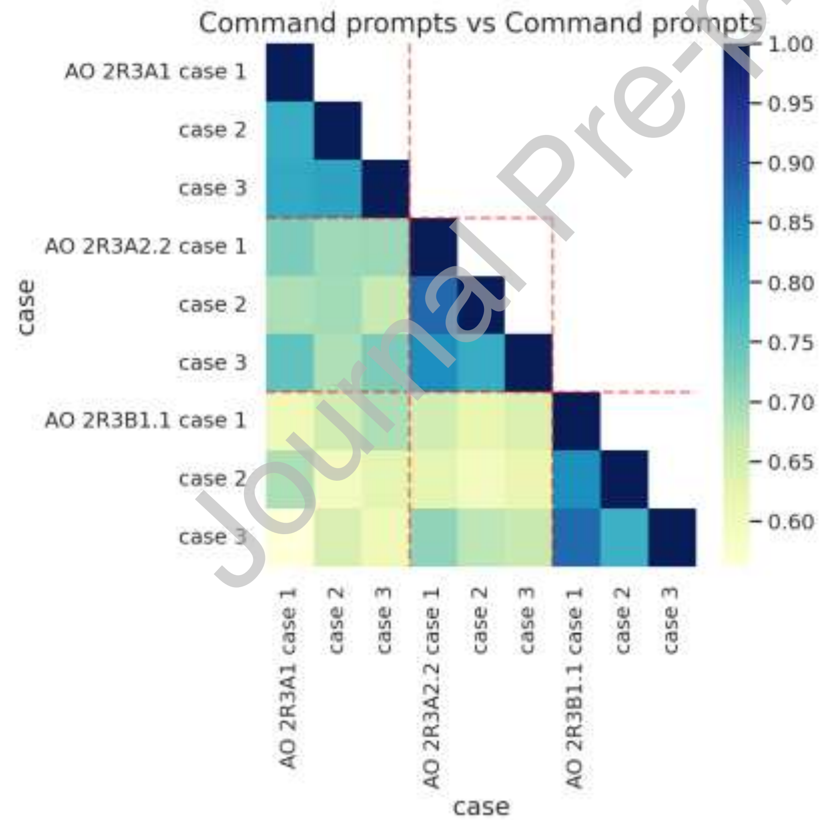


Fig. 1: similarity matrix between command files, computed by bag of words in Python

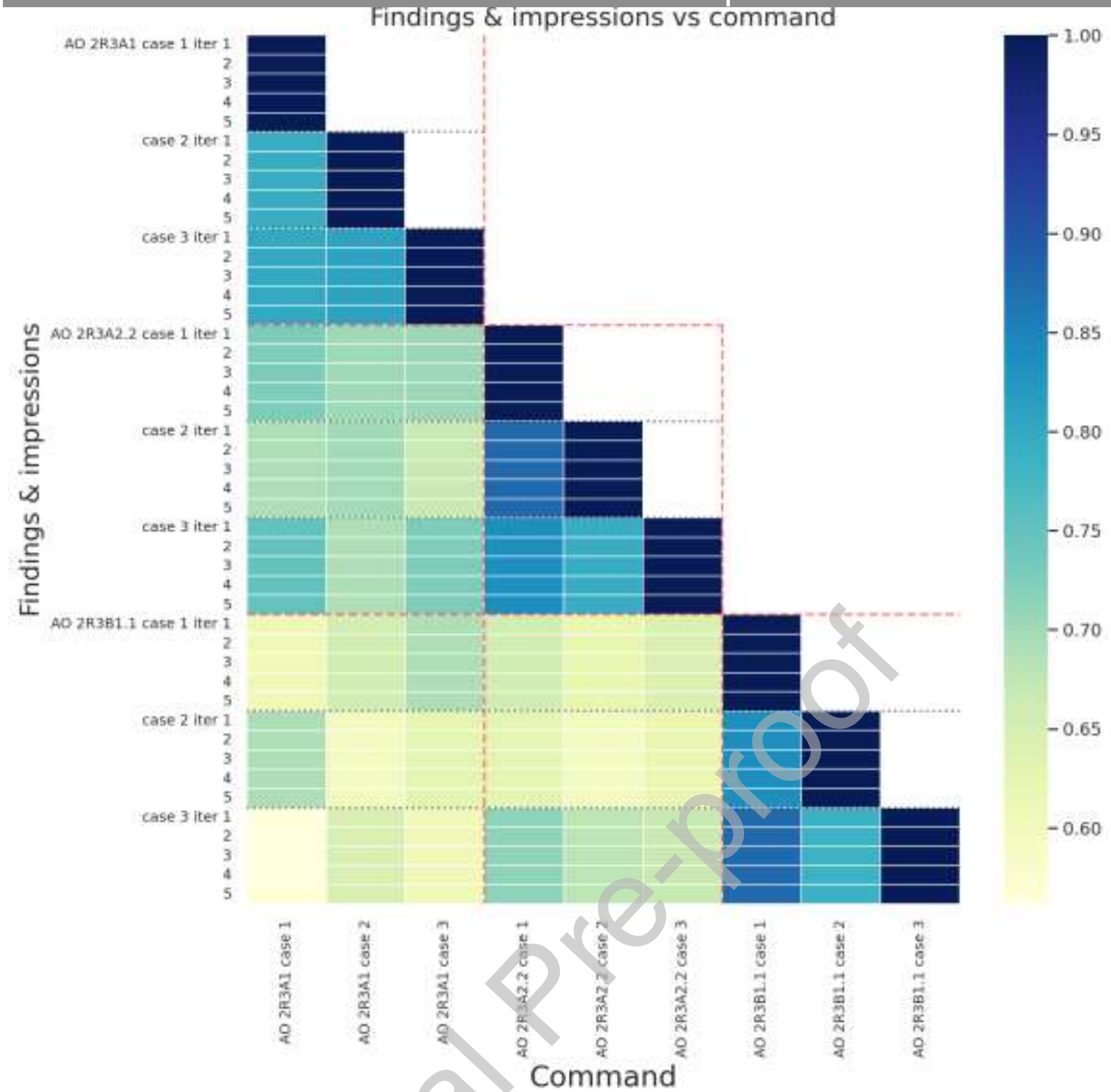


Fig. 2: similarity matrix between command files and returned radiology report drafts under the style setting of findings / impression separately, computed by bag of words in Python

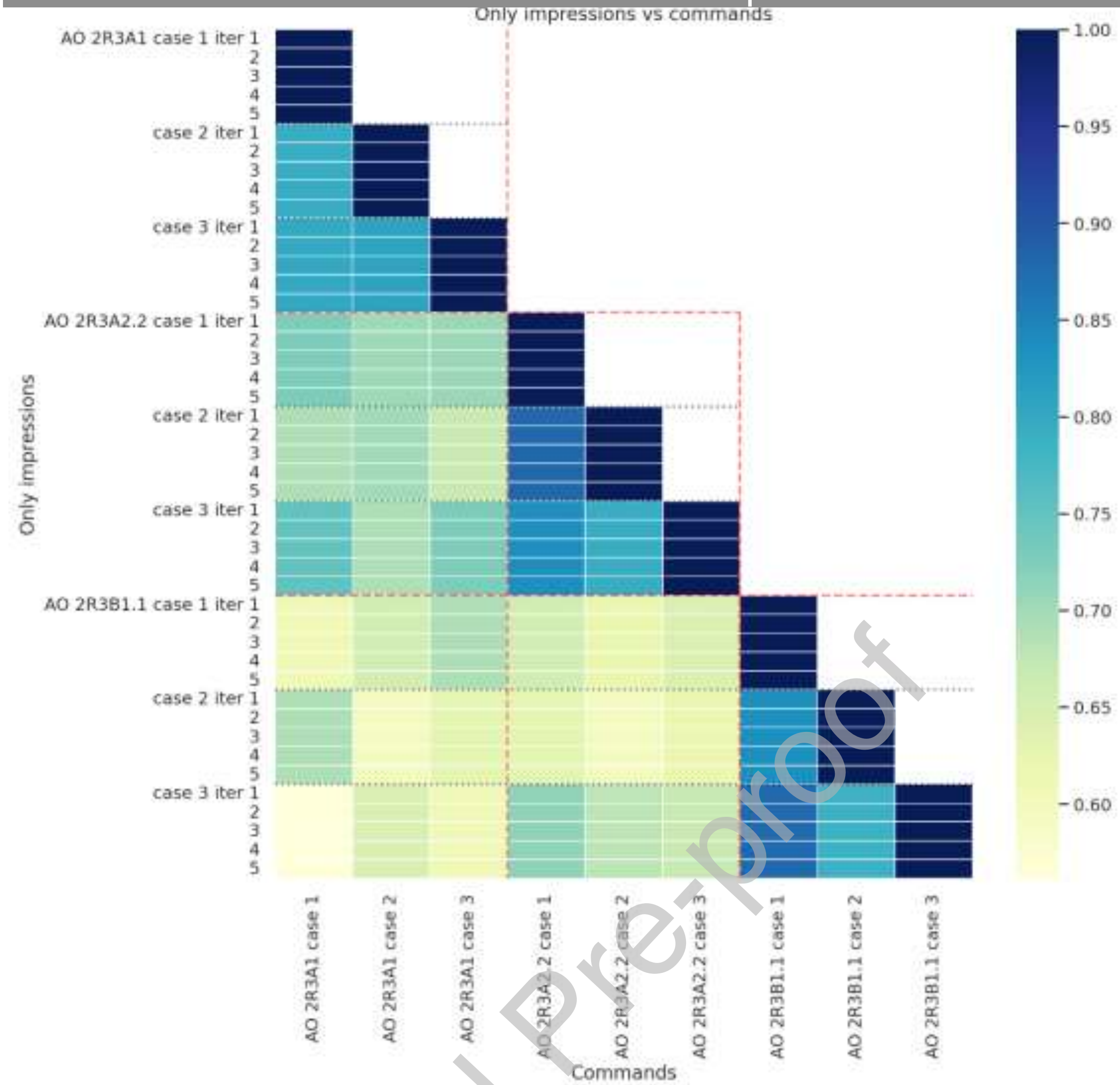


Fig. 3: similarity matrix between command files and returned radiology report drafts under the style setting of findings / impression merged, computed by bag of words in Python

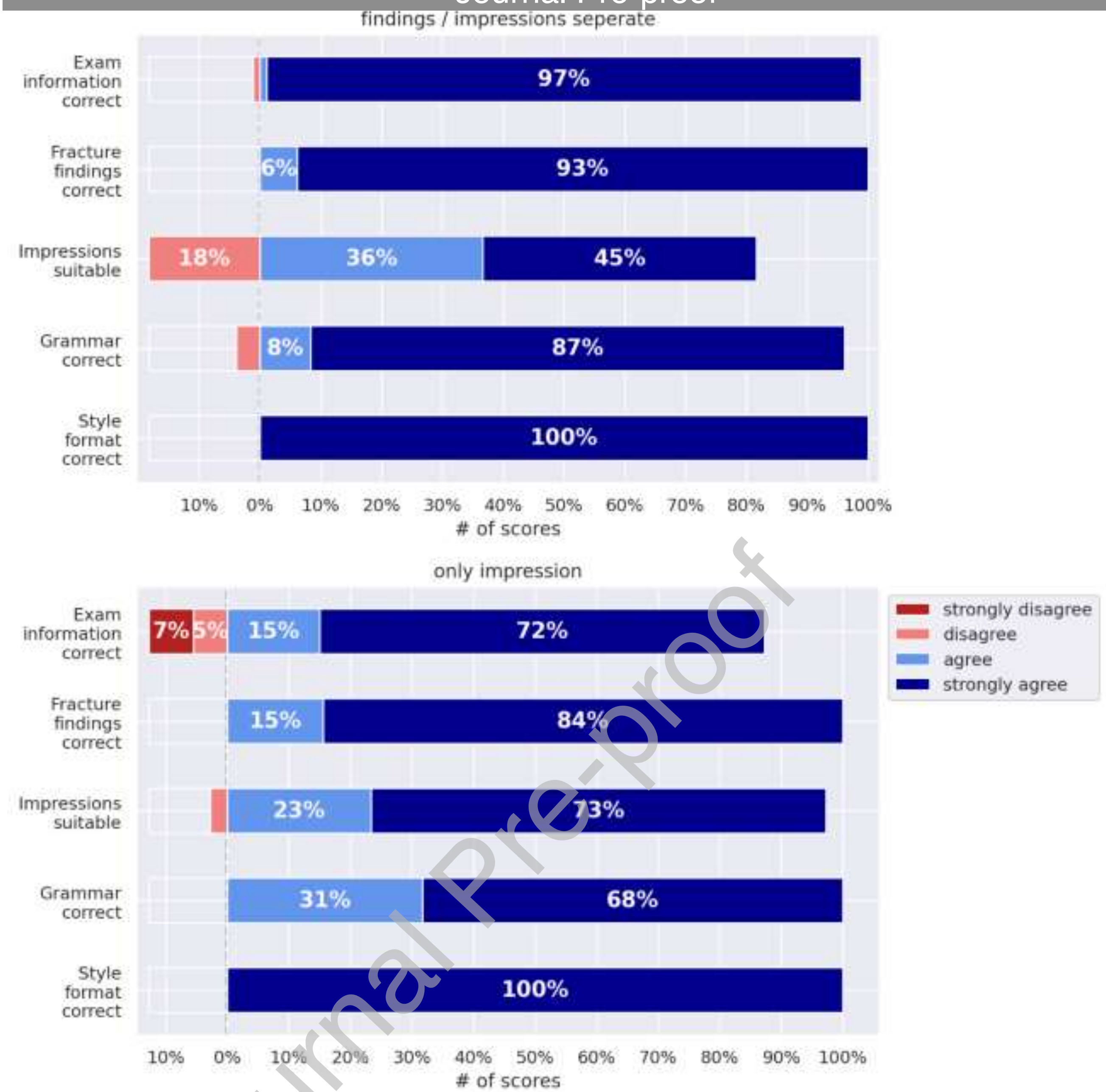


Fig. 4: Score card results from the 4 reviewing radiologists for each of the 5 categories, shown separately for reports with findings / impression separately and merged.