Revista Española
de Educación Médica

Structured radiological report assisted by language models in Radiology residents: pilot implementation in Emergency Department.

Informe radiológico estructurado asistido por modelos de lenguaje en residentes de Radiología: piloto de implementación en Urgencias.

Clemente García-Hidalgo [1*] , José Antonio Consentino Hernández [2] , José Vicente Cayuela Espí [3] , Gonzalo Pagán Vicente [4] , Juana María Plasencia Martínez [5] , Ana Blanco Barrio [6] , Gloria Pérez Hernández [7] , Ana Moreno Pastor [8]

[1] Radiology Department, Morales Meseguer University Hospital, Murcia, Spain. clemente292@gmail.com , https://orcid.org/0009-0001-6672-2714

[2] joseconsentinohernandez@gmail.com ; [3] josevicayuela@gmail.com ; [4] gonzalopv97@gmail.com ; [5] plasen79@gmail.com ; [6] anablancowhite@gmail.com ; [7] gloriapeh93@gmail.com ; [8] anamp78@gmail.com

* Correspondence: Clemente García-Hidalgo clemente292@gmail.com

Received: 3/1/26; Accepted: 20/1/26; Published: 22/1/26

Summary

Objective: To evaluate whether a structured reporting system assisted by Large Language Models (LLMs) can be practically integrated into the work of radiology residents during on-call shifts. Secondary objectives included: describing formatting preferences through blind evaluation, characterizing linguistic differences between manual and LLM-assisted reports, and identifying perceived risks for a confirmatory study. **Methods:** A two-component pilot study was conducted. In the prospective phase, four residents generated 480 reports, alternating between manual and LLM-assisted writing (Custom GPT-4o). In parallel, 200 anonymized reports from attending physicians were analyzed to contextualize the metrics. An ad hoc Likert-type survey (six dimensions) was used, and classification and perplexity metrics were calculated as descriptive indicators. **Results:** The tool was well received. Median Likert scores ranged from 4.75 to 4.90 out of 5. Residents accurately distinguished which reports had been assisted (F1 = 0.92), suggesting a recognizable formal signature. Self-attribution bias was observed in blinded preferences. Perplexity differed between residents and attending physicians ($p = 0.03$), suggesting greater regularity among experienced professionals. **Conclusions:** The findings support the initial integration of the assistant into the on-call system. The value lies in their role as a scaffold to standardize communication between residents and requesting physicians, not in automating diagnostic reasoning.

Keywords: Artificial intelligence; Large language models; Structured report; Medical education; Radiology; Residents; Emergency radiology; Pilot study

Resumen

Objetivo: Evaluar si un sistema de informe estructurado asistido por Large Language Models puede integrarse de forma práctica en el trabajo de residentes de Radiología durante las guardias. Como objetivos secundarios: describir preferencias de formato mediante evaluación ciega, caracterizar diferencias lingüísticas entre informes manuales y asistidos, e identificar riesgos percibidos para un estudio confirmatorio. **Métodos:** Estudio piloto con dos componentes. En la fase prospectiva, cuatro residentes generaron 480 informes alternando redacción manual y asistida por

LLM (Custom GPT-4o). En paralelo, se analizaron 200 informes anonimizados de adjuntos para contextualizar las métricas. Se empleó una encuesta ad hoc tipo Likert (seis dimensiones) y se calcularon métricas de clasificación y perplejidad como indicadores descriptivos. **Resultados:** La herramienta fue bien recibida. Las medianas Likert oscilaron entre 4,75 y 4,90 sobre 5. Los residentes distinguieron con precisión qué informes habían sido asistidos (F1 = 0,92), lo que sugiere una huella formal reconocible. Se observó sesgo de autoatribución en las preferencias ciegas. La perplejidad difirió entre residentes y adjuntos (p = 0,03), apuntando a mayor regularidad en profesionales experimentados. **Conclusiones:** Los hallazgos respaldan la integración inicial del asistente en el circuito de guardias. El interés reside en su función de andamiaje para estandarizar la comunicación entre residentes y médicos peticionarios, no en automatizar el razonamiento diagnóstico.

Palabras clave: Inteligencia artificial; Grandes modelos de lenguaje; Informe estructurado; Educación médica; Radiología; Residentes; Radiología de urgencias; Estudio piloto

1. Introduction

Artificial intelligence is transforming contemporary radiological practice (1-2), with applications that extend beyond interpretation. Among these, structured reporting assisted by large language models (LLM) (3-4) represents an opportunity to improve radiological communication, especially in emergency settings, where speed and accuracy are critical. The radiological report is the primary communicative product of the radiological procedure. From a training perspective, it is an observable outcome of clinical communication skills. However, explicit instruction in its writing is inconsistent. Many residents learn through exposure, imitation, and variable feedback (5). This gap justifies interventions that provide structure, explicit criteria, and feedback during practice. Beyond individual stylistic differences, the report's structure encodes the reasoning underlying the interpretation. Emergency radiology presents characteristics that make standardization relevant: rapid communication, immediate decision-making, and terminological consistency.

In medical education, report quality is not merely a clinical product; it constitutes evidence of clinical reasoning and professional communication. The literature suggests that structured reports can be used to support competency-based training, allowing for the identification of milestones and achievements through observable criteria (6). They can also reduce omissions and improve clinician satisfaction (7). Our study evaluates whether an LLM-based assistant can act as cognitive scaffolding for residents, and what objective linguistic cues emerge from its use.

We frame this intervention as on-the-job learning. The educational mechanism operates through scaffolding (a stable script of sections reduces extrinsic cognitive load), deliberate practice (repetition with explicit criteria), and immediate feedback (style cues that the resident compares with the attending physician's review). The product, the report, thus becomes evidence of communicative competence, without the system being expected to perform diagnostic reasoning.

This study is designed as an implementation investigation in real-world on-call shifts. The challenge is not whether the model generates text, but whether it can be safely integrated into the resident's work. The objective is to assess whether the integration is practical and acceptable, and to identify risks to guide a subsequent trial. The incorporation of generative AI into services necessitates the inclusion of critical skills for residency training: verification, bias management, and secure communication. Evaluating early implementations provides evidence for designing this training. The primary objective was to assess whether the assistant can be integrated into the practice of emergency medicine residents. The secondary, exploratory objectives were to describe formatting preferences, characterize linguistic differences between manual and assisted reports, and identify risks for a confirmatory study.

2. Methods

2.1 Evaluation framework

The study prioritized integration into real-world practice and the identification of barriers before assessing effectiveness. This approach is consistent with the sequence in implementation science: pilot studies as a prerequisite for confirmatory studies. The outcomes correspond primarily to Kirkpatrick level 1 (reaction) and, to a limited extent, to level 2 (perceived learning). By design, we avoided making claims about behavioral changes or impact on care.

2.2 Study design

This pilot study had two components. In the prospective phase, residents generated reports, alternating between manual and LLM-assisted writing. In the retrospective analysis, an independent corpus of anonymized attending physician reports was used to contextualize metrics. Given its pilot nature, the study describes feasibility and generates hypotheses, without establishing causal relationships. Although the corpus includes multiple reports per author, the effective sample size (n) is determined by the number of authors (four residents and the attending physicians). Statistical comparisons are interpreted as exploratory.

2.3 Ethical considerations and safety

The reports were anonymized through a double-checking process: first, automatic removal of direct and quasi-identifiers using regular expressions (names, medical record numbers, dates of birth, addresses); second, manual review to detect potentially identifying contextual references. The tool was used as a writing assistant; diagnostic responsibility remained with the resident and attending physician. Interaction with the LLM excluded identifying patient data. Operational measures included: exclusive use as a writing aid, pre-validation review, and a prohibition on entering identifiers. The assistant does not interpret images or enter findings; it structures the text entered by the resident.

Responsibility for the final content remained with the professional, consistent with the principle of human oversight (8). The usage parameters and safeguards for the assistant were made explicit, following emerging recommendations for studies with generative models, particularly the principles of transparency, human oversight, and delimitation of the system's role proposed in TRIPOD-LLM, as well as the applicable items from the CLAIM checklist (9). Specifically, the non-diagnostic purpose of the model was defined, its configuration and context of use were described, data anonymization was ensured, clinical responsibility was maintained with the human professional, and any automation of diagnostic decision-making was avoided.

This study is an educational and implementation evaluation, without clinical intervention or modification of patient care. All reports were anonymized before analysis. In accordance with Royal Decree 957/2020 and Spanish Law 14/2007 on Biomedical Research, studies using fully anonymized data and not involving patient intervention are exempt from ethics committee approval. The work was conducted in accordance with the principles of the Declaration of Helsinki.

2.4 Tool Development

A custom tool based on Custom GPT-4o (OpenAI, May 2024 version) was developed, designed using prompt engineering without programming. The system was configured with the following parameters: temperature 0.3 (to prioritize consistency over creativity), max_tokens 2048, and no fine-tuning. Reference documentation included the European Society of Emergency Radiology Guidelines (10), RadReport templates from the Radiological Society of North America (11), and the Spanish Society of Medical Radiology's document on conflicting terminology (12).

The implemented functionalities were: automatic detection of modality and anatomical region, contextual suggestions during writing, reminders to reduce omissions, improved speech recognition, and augmented generation with reference documentation (RAG) (13). The structure of the generated structured report followed five standardized sections: Technique (modality, contrast, protocol), Findings (systematic description by anatomical region), Comparison (with previous studies when available), Diagnostic Impression, and Recommendations. Table 1 details the report structure and its correspondence with the assessed communicative competencies.

Table 1. Structure of the structured report and associated communication skills.

Report section	Content	Communicative competence
Technique	Modality, contrast, protocol	Terminological precision
Findings	Systematic description by region	Completeness, organization
Comparison	Changes compared to previous studies	Longitudinal integration
Diagnostic impression	Synthesis and differential diagnosis	Clinical reasoning
Recommendations	Follow-up, additional studies	Clinical orientation

2.5 Population and procedure

Four radiology residents (three second-year residents and one fourth-year resident) participated between June and December 2025. The sample size was determined by operational feasibility: it corresponded to the residents in the department who agreed to participate voluntarily in the pilot study (4 out of 11 total residents, 36%). This participation rate is reasonable for an intervention requiring additional commitment during on-call shifts. The sample size is consistent with recommendations for feasibility pilot studies, where $n \geq 12$ observations per condition is considered sufficient to estimate variability (14). Each resident generated 120 reports (60 assisted, 60 manual), totaling 480. Cases were assigned consecutively, alternating the reporting method. No formal counterbalancing was established, which is a recognized limitation. Table 2 summarizes the characteristics of the participants and the distribution of reports.

Table 2. Characteristics of participants and distribution of reports.

Feature	Res. A	Res. B	Res. C	Res. D	Total
Year of residence	R2	R2	R2	R4	-
Assisted reports	60	60	60	60	240
Manual reports	60	60	60	60	240
Total per resident	120	120	120	120	480
Modalities	CT scan (65%), X-ray (30%), Ultrasound (5%)				

2.6 Training intervention

A brief (30-minute) induction was provided, focusing on the educational objective (improving structure, completeness, and clarity), safe use, and quality criteria for emergency department reports. The tool was used during the report writing process; subsequently, the report followed the standard review process by the on-call attending radiologist.

2.7 Multidimensional evaluation

A Likert-scale survey (1-5) was designed to assess six dimensions: Usability, Clinical Utility, Efficiency, Reliability, Perceived Educational Impact, and Overall Satisfaction. The items evaluate communicative quality, not clinical detection capabilities. The survey was developed through consensus among the research team, reviewing similar instruments in the health technology implementation literature (15), although without formal psychometric validation given the exploratory nature of the study. A preliminary test was conducted with two non-participating residents to verify the comprehensibility of the items. The survey was conceived as formative feedback to capture early signals. Given its pilot nature, psychometric validity is not assumed; the results are exploratory in nature and should be confirmed with validated instruments in subsequent studies.

2.8 Qualifying phase

To explore whether the LLM leaves a perceptible formal imprint, a blind identification task (assisted vs. manual) was performed on the 480 reports. The reports were preprocessed by removing headings and signatures and presented in randomized order using dedicated software. Each evaluator only classified reports from other authors to avoid recognition bias. Accuracy, precision, sensitivity, and F1 were calculated as descriptive metrics of human discriminatory ability, not as indicators of differential quality.

2.9 Blind preference

A blinded preference experiment was designed to assess attribution bias. For each clinical case with two reports (assisted and manual), anonymized pairs were generated and presented in random order. Identifiable markers of authorship and method were removed. Each evaluator assessed only pairs of reports generated by other residents, never their own, to avoid stylistic recognition. Participants were instructed to choose their preferred report based solely on clarity, completeness, and organization, without knowledge of the method used to generate each report. This component was exploratory, aimed at detecting potential attribution bias rather than establishing the superiority of one method.

2.10 Analysis of perplexity

Two hundred reports from attending radiologists from 2025 were retrieved, anonymized, and normalized. Perplexity was used as an exploratory metric of linguistic regularity, not as an educational outcome or indicator of clinical quality. The calculation was performed using the OpenAI API (gpt-4-1106-preview). It is important to note that lower perplexity only indicates greater predictability of the text for the language model used, without necessarily implying better communicative or diagnostic quality. The values between resident (n = 480) and attending (n = 200) reports were compared using the Mann-Whitney U test, a non-parametric test appropriate for non-normally distributed data. Perplexity was calculated as "Perplexity" = exp {left (- {1 } over {N } sum from {i=1 } to {N } {log {p} left ({w} rsub {i} right) } right)} , "where " p left ({w} rsub {i} right) " is the probability assigned by the model to the token " i.

2.11 Statistical analysis

The results are presented descriptively. Likert scales are summarized using the median and interquartile range, appropriate for ordinal data. No parametric tests were used. Computational metrics are interpreted as exploratory analyses. The analyses were performed using Python 3.11 (pandas, scipy, and scikit-learn libraries).

3. Results

3.1 Acceptability Assessment

Table 3 shows the Likert scale results. All dimensions had medians above 4.7 out of 5. Figure 1 graphically presents these results. Overall Satisfaction had the highest median (4.90), followed by Usability and Efficiency (4.85). The lowest was Reliability (4.75), which could reflect residents' appropriate caution regarding the need for verification of the generated content.

3.2 Classification Analysis

Table 4 shows the metrics from the linguistic fingerprinting experiment, where human evaluators classified the reports. The F1 score of 0.92 indicates that the human evaluators were able to distinguish with high accuracy between assisted and manual reports. This suggests distinguishable linguistic or structural features, although these differences are formal and do not imply differences in clinical quality or diagnostic utility. The ability to distinguish between methods could be due to greater structural consistency, more systematic use of standardized terminology, or organizational patterns characteristic of the assistant-assisted method.

Table 3. Acceptability assessment of the LLM-assisted structured report.

Dimension	Eval. A	Evaluation B	Eval. C	Eval. D	Median (IQR)
Usability	4.8	4.9	4.7	5.0	4.85 (4.7-5.0)
Clinical Utility	4.9	4.7	4.8	4.9	4.80 (4.7-4.9)
Efficiency	4.7	4.8	4.9	4.9	4.85 (4.7-4.9)
Reliability	4.6	4.8	4.7	4.9	4.75 (4.6-4.9)
Educational Impact	4.8	4.9	4.6	4.8	4.80 (4.6-4.9)
Overall Satisfaction	5.0	4.8	4.9	4.9	4.90 (4.8-5.0)

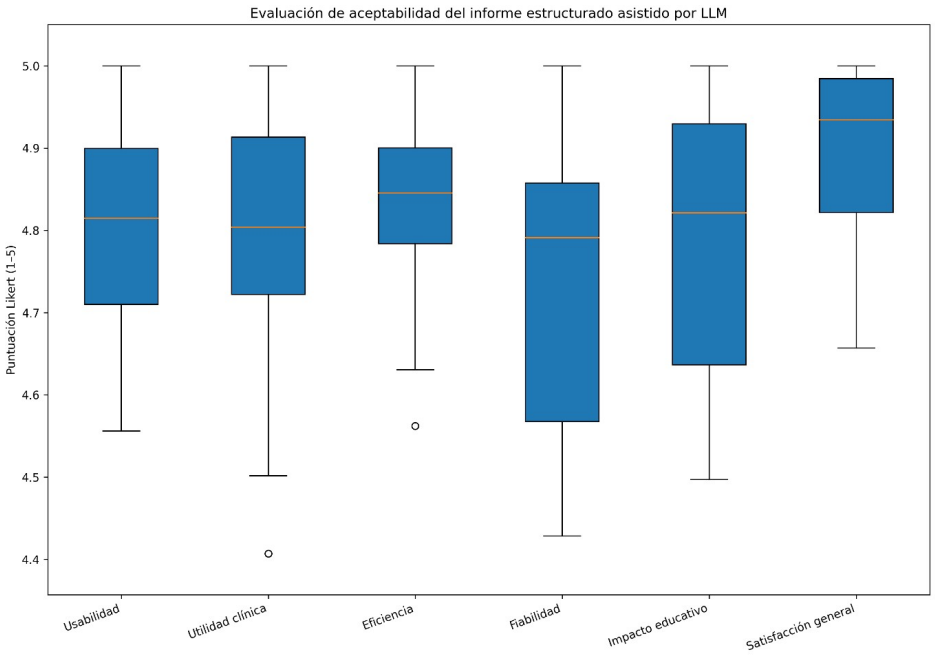

Figure 1. Box and whisker plot with Likert medians and interquartile ranges by dimension. Y-axis: score 1-5, with adjusted scale; X-axis: dimensions evaluated.

Table 4. Exploratory classifier metrics (human evaluators).

Metrics	Worth
Accuracy	0.92
Precision	0.91
Sensitivity	0.94
Specificity	0.91
F1-score	0.92

3.3 Blind preference

A systematic difference was observed according to the relationship with the author: evaluators more frequently selected reports from colleagues (78%) than their own (70%). This pattern suggests possible attribution bias, although interpretation should be cautious given the limited sample size.

3.4 Perplexity

Resident reports showed greater perplexity (median 28.5; IQR 25.0–30.5) than those of attending physicians (median 26.0; IQR 24.0–29.0; p = 0.03). Figure 2 illustrates this comparison. This exploratory finding suggests greater linguistic regularity in experienced professionals, possibly reflecting more standardized writing patterns acquired with experience. It is important to reiterate that perplexity measures textual predictability for a specific language model, not the clinical quality of the report.

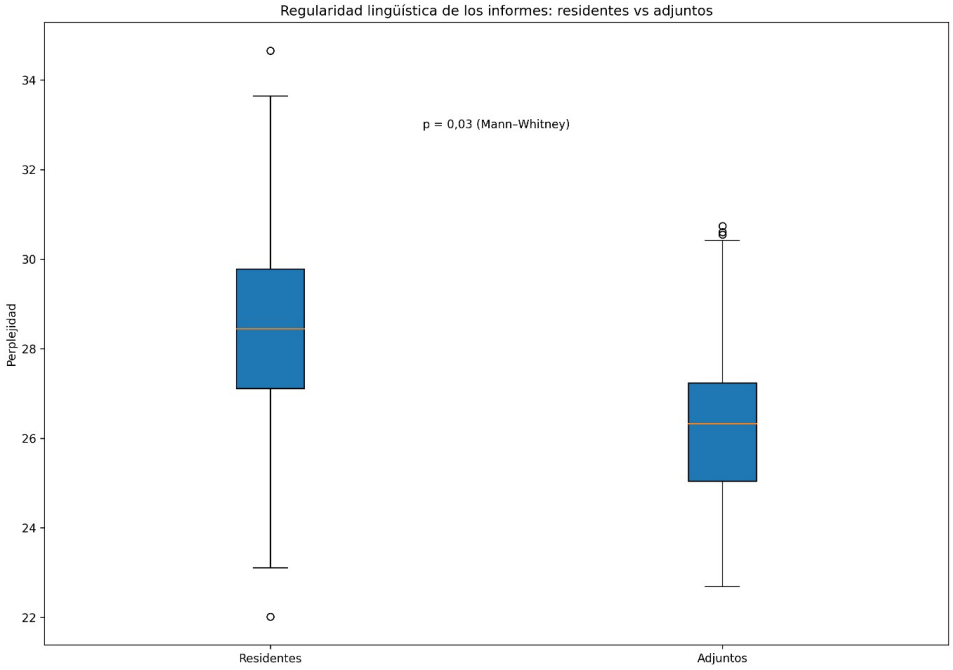

Figure 2. Box and whisker plot comparing perplexity between resident and attending physician reports.

4. Discussion

This pilot study explores the implementation of an LLM-assisted structured reporting tool in the emergency department, developed by residents without advanced programming. The results provide initial favorable signals regarding feasibility and acceptance, although design limitations preclude establishing causal relationships or generalizing the findings.

4.1 Interpretation of results

The high median Likert scores (>4.7/5) reflect good initial acceptance. Overall Satisfaction reached 4.90, suggesting that the experience met participants' expectations. Usability (4.85) indicates that integration was perceived as seamless into the on-call workflow. Reliability, with the lowest median (4.75), can be interpreted favorably as reflecting an appropriate critical attitude, where residents recognize that the final diagnostic responsibility rests with the radiologist and not the tool. This underscores the importance of framing LLMs as support tools (16), not as autonomous clinical decision-making systems.

4.2 Connection between results and training impact

The results obtained can be interpreted in relation to the proposed educational impact. The high acceptability (medians >4.7) suggests that the tool does not generate resistance or interrupt the learning process, a necessary condition for any training intervention. The lower level of perplexity observed in attending physicians compared to residents is consistent with the hypothesis that professional experience leads to more standardized communication patterns; if the tool facilitates the early adoption of these patterns, it could accelerate the learning curve, although this hypothesis requires longitudinal confirmation. The residents' ability to distinguish assisted reports (F1 = 0.92) indicates that they recognize the characteristics of the structured format, a step prior to its internalization. Taken together, these indicators are compatible with the scaffolding theoretical framework: recognizable external support that can be progressively withdrawn as skills are consolidated (17).

4.3 Linguistic footprint

The F1-score of 0.92 on the human classification task demonstrates that there are distinguishable formal characteristics between assisted and manual reports. However, this finding should be interpreted with caution: it indicates a perceptible stylistic difference, but does not imply value judgments about which modality produces reports of higher clinical or communicative quality. The difference in perplexity between residents and attending physicians suggests that professional experience leads to greater standardization of writing style, a finding consistent with the literature on expertise development (cf. 7,20).

4.4 Educational Implications

The potential educational value of this tool lies not in automating the report, but in developing transversal skills: organization of reasoning, systematic completeness, and communicative clarity. The concept of cognitive scaffolding, borrowed from the sociocultural theory of learning (17), suggests that temporary support structures can facilitate the development of skills that are subsequently internalized. In this case, the structured template acts as an external organizer that reduces extrinsic cognitive load, allowing the resident to concentrate on the diagnostic content.

As a concrete example, the assistant includes systematic reminders to verify the description of findings in all relevant anatomical regions, reducing inadvertent omissions. It also suggests including comparisons with previous studies when available, an element frequently overlooked in emergency department reports. These mechanisms illustrate how scaffolding can translate into tangible improvements in completeness, regardless of the diagnosis.

This pilot only assesses operational integration; demonstrating genuine learning will require external measures and longitudinal studies. In competency-based training, standardization can facilitate the formative assessment of observable components of reasoning, aligned with EPAs and competency progression (6). The medical education literature recognizes the value of observable

products such as the structured report to document communicative competencies and serve as a basis for targeted feedback (18-19).

In a confirmatory study, the impact should be measured with observable outcomes: report completeness (percentage of sections covered), terminological consistency (adherence to standardized lexicon), preparation time, and resident-attending disagreement rate.

4.5 Context in literature

Recent reviews highlight that, despite the potential of structured reporting, its adoption in clinical practice remains limited. LLMs are proposed as catalysts to overcome implementation barriers, although regulatory and validation challenges persist (21). A natural extension of this work would be to evaluate discrepancies between generated drafts and validated final versions, as proposed by studies that categorize textual, semantic, and critical changes in radiology reports (22).

4.6 Decentralized Development

A distinctive feature is that the tool was developed by residents using prompt engineering, requiring no programming knowledge. This decentralized development model democratizes access to AI technologies, although it raises validation and governance challenges that will need to be addressed as similar tools proliferate in clinical settings.

4.7 Educational Risks

Generative assistants can produce immediate operational benefits, but they can also displace the cognitive exercise of describing and writing. The potential for efficiency must be weighed against the risk of technological dependence, analogous to that described for GPS navigation systems (23). In radiology, this risk could affect the resident's core competencies if the tool is used as a substitute for diagnostic thinking rather than as temporary scaffolding. We propose a pedagogically safe use with three safeguards: that the resident formulate the diagnostic impression in their own words before using the tool, that the attending physician provide feedback focused on structure and reasoning, and that a gradual withdrawal of support is applied to advanced residents.

4.8 Strengths and limitations

This study provides an implementation in real-world emergency settings, with high ecological validity. The report stands as observable formative evidence, and generative AI is framed as non-diagnostic support integrated into routine supervision. The limitations are substantial and must be considered when interpreting the results. The sample size (four volunteer residents out of a possible eleven, without a control group) limits generalizability and statistical power. The use of an ad hoc survey not psychometrically validated introduces uncertainty about the reliability and validity of the measures. The perceptual nature of the "educational impact" does not allow for asserting actual learning without longitudinal evaluation using external measures. The carry-over risks inherent in the alternating design without formal counterbalancing may have introduced sequencing bias. Finally, the reliance on a human classifier to detect the linguistic footprint introduces subjectivity. The results constitute preliminary evidence that justifies, but does not replace, multicenter controlled studies with adequate samples and validated instruments.

5. Conclusions

- In light of this pilot, the findings suggest that the initial integration of LLM-assisted structured reporting in radiodiagnostic residents is feasible and well-accepted in the context of on-call shifts.
- From the perspective of medical education, the interest of this tool lies not in automating diagnostic reasoning, but in its potential as a scaffold that promotes more standardized communication, reduces formal variability, and can turn the report into observable evidence for competency-based training frameworks (18, 20, 24).
- As a preliminary finding, these results justify further studies with controlled designs, adequate samples, and external measures focused on communicative performance and objective report quality.

Funding: There has been no funding.

Declaration of conflict of interest: The authors declare that they have no conflict of interest.

Authors' contributions: All authors contributed to the conceptualization, methodology, data acquisition, formal analysis, research, drafting of the original manuscript, revision and editing of the manuscript, and approval of the final version.

6. References

1. Kahn CE Jr. Artificial intelligence in radiology: decision support systems. *Radiographics*. **1994**, 14, 849-861. https://doi.org/10.1148/radiographics.14.4.7938772
2. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. *N Engl J Med*. **2023**, 388, 1981-1990. https://doi.org/10.1056/NEJMra2301725
3. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*. **2020**, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165
4. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*. **2017**, 30, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762
5. Hartung MP, Bickle IC, Gaillard F, Kanne JP. How to Create a Great Radiology Report. *RadioGraphics*. **2020**, 40, 1658-1670. https://doi.org/10.1148/rg.2020200020
6. Castro D, Mishra S, Kwan BY, et al. Structured Reporting in Radiology Residency: A Standardized Approach to Assessing Interpretation Skills and Competence. *Int Med Educ*. **2025**, 4, 40. https://doi.org/10.3390/ime4010002
7. Larson DB, Towbin AJ, Pryor RM, Donnelly LF. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. *Radiology*. **2013**, 267, 240-250. https://doi.org/10.1148/radiol.12121502
8. Kao JP, Kao HT. Large Language Models in radiology: A technical and clinical perspective. *Eur J Radiol Artif Intell*. **2025**, 2, 100021. https://doi.org/10.1016/j.ejrai.2025.100021
9. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell*. **2020**, 2, e200029. https://doi.org/10.1148/ryai.2020200029
10. Wirth S, Hebebrand J, Basilico R, et al. European Society of Emergency Radiology (ESER). Guideline on radiological polytrauma imaging and service (full version). Available in: https://www.eser-society.org/polytrauma-imaging-guidelines/ (Accessed: January 2025).
11. Radiological Society of North America. RadReportTemplates. Available in: https://www.rsna.org/practice-tools/data-tools-and-standards/radreport-templates (Accessed: January 2025).
12. Spanish Society of Medical Radiology. *Conflicting Lexicon in Radiology* . Madrid: SERAM; 2020. Available from: https://static.seram.es/wp-content/uploads/2021/07/lexico_radiologico_conflictivo.pdf (Accessed: January 2025).
13. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*. **2020**, 33, 9459-9474. https://doi.org/10.48550/arXiv.2005.11401
14. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract*. **2004**, 10, 307-312. https://doi.org/10.1111/j..2002.384.doc.x

15. Brooke J. SUS: A 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, eds. *Usability Evaluation in Industry* . London: Taylor & Francis; **1996**. p. 189-194.
16. Patel BN, Rosenberg L, Willcox G, et al. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med*. **2019**, 2, 111. https://doi.org/10.1038/s41746-019-0189-7
17. Wood D, Bruner JS, Ross G. The role of tutoring in problem solving. *J Child Psychol Psychiatry*. **1976**, 17, 89-100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x
18. Ten Cate O. Entrustability of professional activities and competence-based training. *Med Educ*. **2005**, 39, 1176-1177. https://doi.org/10.1111/j.1365-2929.2005.02341.x
19. Epstein RM. Assessment in medical education. *N Engl J Med*. **2007**, 356, 387-396. https://doi.org/10.1056/NEJMra054784
20. Schwartz LH, Panicek DM, Berk AR, et al. Improving communication of diagnostic radiology findings through structured reporting. *Radiology*. **2011**, 260, 174-181. https://doi.org/10.1148/radiol.11101913
21. Busch F, Hoffmann L, Pinto dos Santos D, et al. Large language models for structured reporting in radiology: past, present, and future. *Eur Radiol*. **2025**, 35, 2589-2602. https://doi.org/10.1007/s00330-024-11107-6
22. Lindholz M, Burdenski A, Ruppel R, et al. Comparing large language models and text embedding models for automated classification of textual, semantic, and critical changes in radiology reports. *Eur J Radiol*. **2025**, 191, 112316. https://doi.org/10.1016/j.ejrad.2025.112316
23. Martín-Noguerol T, López-Úbeda P, Luna A. From GPS to ChatGPT in Radiology... Dumb and Dumber? *J Am Coll Radiol*. **2025**. https://doi.org/10.1016/j.jacr.2025.09.014
24. European Society of Radiology. ESR paper on structured reporting in radiology. *Insights Imaging*. **2018**, 9, 1-7. https://doi.org/10.1007/s13244-017-0588-8