Large Language Models for Structured Chest CT Reporting in Portuguese: A Comparative Study with Radiologist Validation

Juliana Petruceli[1], **Marcelo Oliveira**[1], **Tarcisio Ferreira**[1], **Jose Arthur Sabino**[1]

[1] Computing Institute, Federal University of Alagoas (UFAL)

{jsp,oliveiramc,tlf,jals}@ic.ufal.br

Abstract. *To evaluate LLMs for converting free-text Portuguese chest CT reports into structured JSON for clinical communication and data reuse. Gemini 1.5 Flash, GPT-4o, and LLaMA 3.3 were tested on 1,102 de-identified reports using a dynamic JSON template with few-shot prompting. Validation combined radiologist review and quantitative metrics. All models produced coherent structured outputs. Gemini achieved the best agreement (macro-F1 0.852; micro-F1 0.853), followed by LLaMA (0.806; 0.809) and GPT-4o (0.797;0.798). LLM-assisted structuring of Portuguese chest CT reports is feasible and attains high agreement with manual references; section-aware prompting and JSON validation improve robustness.*

1. Introduction

Radiology is one of the primary sources of clinical information in hospitals, and the radiology report is, in practice, the artifact that connects image interpretation to therapeutic decision-making [Pesapane et al. 2023, Goldberg-Stein and Chernyak 2019]. Nevertheless, free-text reporting still predominates, leading to substantial variability in terminology, ordering of findings, and descriptive granularity. Such heterogeneity compromises not only readability for the care team but also the computational reuse of report content—for example, for quality auditing, clinical surveillance, research, and integration with electronic health records and analytical pipelines [McFarland et al. 2021].

Structured reporting initiatives have sought to mitigate these limitations through templates and terminological standardization, such as models promoted by international organizations (e.g., the RSNA) [Pesapane et al. 2023, Nobel et al. 2022]. In parallel, advances in natural language processing (NLP) and deep learning have enabled approaches to convert unstructured reports into normalized formats [Spandorfer et al. 2019, Mozayan et al. 2021, Elvas et al. 2025]. With the consolidation of Transformer architectures, Large Language Models (LLMs) have become capable of performing complex text transformation tasks with reduced reliance on task-specific supervised training, supporting strategies such as zero-shot and few-shot learning [Elvas et al. 2025, Bhayana 2024, Fink et al. 2023].

However, much of the available evidence and resources remains concentrated in English. This constrains generalization to Portuguese, where terminological particularities, writing styles, and local clinical contexts differ. Moreover, structure that is truly useful for health information systems requires more than "well-organized sections": it demands a *machine-readable* output contract, with fields and hierarchies that enable validation, comparison, and downstream integration with clinical systems (PACS/RIS/EHR) and analytical repositories [Woźnicki et al. 2024, Adams et al. 2023].

In this context, the main goal of this paper is to evaluate, using real-world Portuguese chest computed tomography (CT) reports, the ability of different large language models (LLMs) to transform free-text narratives into structured reports through an adaptable JSON template, measuring fidelity, clarity, and machine-readable structural consistency.

To guide our investigation, we address the following research questions:

1. Can LLMs structure Portuguese radiology reports into JSON with high fidelity and consistency?
2. What trade-offs emerge between proprietary models and locally deployed open-source models regarding quality/quantitative metrics, performance, and computational cost?
3. Which report fields and types exhibit the highest error rates, and how effectively can prompting strategies or few-shot examples reduce these errors?

Highlights of this paper: (i) evaluates LLM-based structuring of free-text Portuguese chest CT reports in real-world data; (ii) introduces an adaptable JSON template that preserves radiologists' terminology while omitting absent fields; (iii) benchmarks proprietary and locally deployable open-source models under the same prompting protocol; and (iv) validates outputs through complementary evidence, combining radiologist qualitative assessment with quantitative agreement against a manually structured reference dataset.

2. Materials and Methods

The study was conducted in six successive stages, each addressing specific aspects of structuring radiology reports using LLMs.

Stage 1 — Data preparation. A total of 1,102 chest computed tomography (CT) reports were used, written by different radiologists and obtained from a single hospital-based dataset intended for academic research. Each report was sequentially numbered (1–1102) to enable control and comparison between the original and the structured versions. The texts were kept in full, considering that the evaluated language models could automatically correct potential spelling and grammatical errors during the structuring process.

The project was approved by the Research Ethics Committee of the (blind) (CAAE blind). All reports were anonymized and blinded, ensuring confidentiality in accordance with Brazilian National Health Council Resolution CNS No. 466/2012.

Stage 2 — Development of the structured template. A single, dynamic template was created in JSON format, capable of automatically adapting to the content of each report [Woźnicki et al. 2024, Adams et al. 2023]. The structure included the main components of the examination—exam type, clinical indication, technique, findings (grouped by anatomical systems), diagnostic impression, additional findings, and notes, based on the principles of structured reporting [Goldberg-Stein and Chernyak 2019]. Each field included instructions and examples for completion: mandatory fields (`exam_type` and `technique`) had to be filled in all cases, while optional fields (`indication`, `findings`) and subfields (`additional_findings`, and `note`) were included only when present in the original report (Figure 1). The models were instructed to omit empty fields, ensuring clarity, hierarchical coherence, and fidelity to clinical logic.

```
{
  "tipo_exame": "Tomografia Computadorizada do Tórax",
  "indicacao": "",
  "tecnica": {
      "descricao": "Aquisição volumétrica de imagens em equipamento de tomografia computadorizada multislice, com reformatações multiplanares e reconstruções de alta resolução."
  },
  "achados": {
    "avaliacao_comparativa": "",
    "pulmões_pleura": "",
    "traqueia_bronquios": "",
    "hilos_mediastino": "",
    "coracao_pericardio": "",
    "vasos_mediastinais": "",
    "dispositivos_materiais": "",
    "achados_incidentais": "",
    "parede_toracica": "",
    "elementos_osseos": ""
  },
  "impressao_diagnostica": "",
  "achados_adicionais": "",
  "nota": ""
}
```


Figure 1. Structured template model example in Portuguese (JSON format).

Stage 3 — Prompt engineering and few-shot learning. Based on prior studies in clinical NLP [Fink et al. 2023, Shah 2024, Russe et al. 2024, Dorfner et al. 2024], prompt engineering and the use of few-shot learning examples were essential to guide the models in applying the structured template, ensuring organization and adherence to clinical guidelines. The main prompt guided the restructuring of the original reports according to the JSON template (Figure 1), preserving the integrity of findings and avoiding redundancies or improper additions. Two complex few-shot learning examples were used to teach correct completion under different clinical scenarios.

Stage 4 — Optimization of decoding hyperparameters. To determine the most suitable combination of decoding parameters for the automatic structuring of radiology reports in Portuguese, an exploratory study was conducted using the LLaMA 3.3 model. Eight combinations were tested (Temperature: 0.1 and 0.3; Top-K: 10 and 20; Top-P: 0.1 and 0.3), totaling 168 generated reports from 21 originals. Each output was manually evaluated by the radiologists according to four criteria—fidelity to the original content, compliance with the JSON structure, textual clarity, and spelling correctness—scored from 0 to 2 points per criterion.

Stage 5 — Qualitative validation by radiologists. Qualitative validation involved 12 radiologists—six specialists with more than 10 years of experience and six residents with up to three years of training—ensuring a balance between clinical expertise and textual perception [Tam et al. 2024]. Three language models were evaluated: Gemini 1.5 Flash, GPT-4o, and LLaMA 3.3, representing different architectures and access modalities [Woźnicki et al. 2024, Tam et al. 2024]. The sample included 60 reports (12 normal, 20 comparative, and 28 complex), analyzed in a blinded process with partial cross-review to ensure impartiality [Boateng et al. 2018]. Each radiologist evaluated ten sets consisting of the original text and the structured versions generated by the models. Assessments followed a three-point scale (0 = unsatisfactory, 1 = intermediate, 2 = satisfactory) applied to fidelity, clarity, and structure, in addition to selecting the preferred model.

Results were consolidated into standardized spreadsheets, considering mean scores per criterion and preference frequencies, in line with good practices for qualitative

validation in radiology [Tam et al. 2024].

Stage 6 — Construction of the manual reference for quantitative validation. To enable quantitative analysis, a manual reference in JSON format was created, structured according to the single adaptable template and validated by a radiologist with over 20 years of clinical experience. The reference was based on the same 60 reports evaluated in the previous stage, with faithful transcription of the original texts and minimal spelling corrections, preserving original content without clinical inference. Each reference received standardized naming (`reference_ID.json`), enabling automated comparison with model outputs and computation of precision, recall, and F1-score. This approach ensured compatibility between human and automatic versions and enabled objective measurement of LLM performance in a clinical context [Woźnicki et al. 2024, Dorfner et al. 2024, Tam et al. 2024]. Quantitative evaluation measured model performance in report restructuring using precision, recall, and F1-score—widely used metrics for assessing structuring tasks in NLP applied to radiology [Elvas et al. 2025, Woźnicki et al. 2024, Dorfner et al. 2024, Takita et al. 2025].

3. Results and Discussion

The results of applying large language models (LLMs) to restructure chest CT reports are presented below.

3.1. Decoding hyperparameter optimization results

The evaluation of eight decoding combinations for LLaMA 3.3 showed stable performance, with mean scores ranging from 6.33 ± 1.15 to 6.48 ± 1.08. The Temperature 0.3, Top-P 0.3 and Top-k 20 configuration achieved the best result (6.48 ± 1.08) and was adopted as the standard for subsequent stages.

Combinations with temperature 0.3 showed higher semantic fidelity (1.48 ± 0.05) and better adherence to the JSON template (1.67 ± 0.03), while textual clarity reached 1.76 ± 0.05 in configurations with Top-P 0.3 and Top-K 20. The orthographic criterion remained stable (1.57 ± 0.00).

Recurring failures persisted, omissions in comparative reports, JSON duplications, and preservation of spelling errors, attributed to the need for more specific prompts. Thus, final quality proved to be more influenced by instruction formulation and report complexity than by the hyperparameters themselves.

3.2. Qualitative validation results

Qualitative validation assessed the textual and structural quality of reports restructured by Gemini 1.5 Flash, GPT-4o, and LLaMA 3.3, considering three main criteria: fidelity, clarity, and structure. Mean scores clustered near the upper end of the scale (≈ 2.0), indicating satisfactory performance by all models (Figure 2). GPT-4o showed better consistency and lower variability, followed closely by Gemini 1.5 Flash, while LLaMA 3.3 exhibited slightly lower means and a higher frequency of intermediate scores.

Evaluator preference reinforced this pattern: GPT-4o was most frequently selected (30.1%), followed by Gemini 1.5 Flash (24.4%) and LLaMA 3.3 (12.2%), while 33.3%

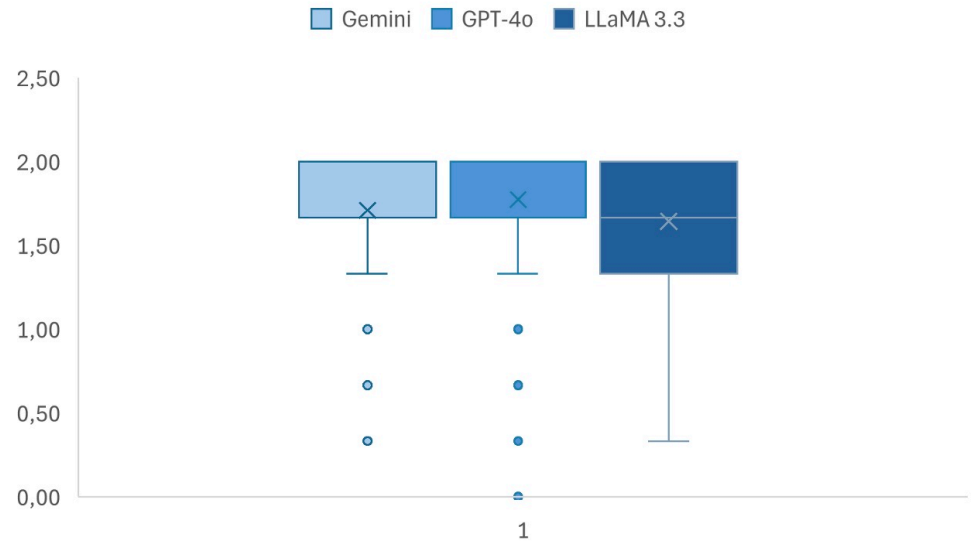

Figura 2. Distribution of the composite score (fidelity, clarity, and structure) in the qualitative evaluation of Gemini 1.5 Flash, GPT-4o, and LLaMA 3.3.

of responses remained neutral. This distribution suggests subtle differences among models, more associated with report complexity than with the intrinsic performance of each system.

It was observed that Gemini 1.5 Flash and LLaMA 3.3 often left empty fields in the JSON template, contrary to the prompt; however, this issue was not penalized by evaluators, which may have slightly overestimated the structural performance of these models.

Qualitative comments indicated difficulty in distinguishing relevant differences between versions, although many radiologists highlighted greater clarity and fluency in texts generated by GPT-4o. Redundancies, empty fields, and occasional omissions were also reported, especially in comparative reports.

In summary, all models demonstrated satisfactory capability to structure reports, with balanced performance between GPT-4o and Gemini 1.5 Flash and lower acceptance of LLaMA 3.3. The observed differences reinforce the need for quantitative complement to objectively measure the structural and textual fidelity of generated outputs.

3.3. Quantitative validation results

Comparing model-structured reports with manual references enabled calculation of the F1-score per report (macro-averages) and in aggregate (micro-averages). Macro-averages represent the simple mean of each report's individual metrics, whereas micro-averages consider the full set of observations, reflecting overall model performance.

Results showed high performance and consistency across evaluated models. Gemini 1.5 Flash achieved the best overall mean performance (F1-score $= 0.852 \pm 0.122$), followed by LLaMA 3.3 (0.806 ± 0.105) and GPT-4o (0.798 ± 0.128). Macro-averages indicated higher structural consistency for Gemini 1.5 Flash, while micro-averages reinforced the stability of LLaMA 3.3 and the satisfactory performance of GPT-4o (Table 1).

These findings confirm the effectiveness of LLMs for automated structuring of radiology reports in Portuguese, demonstrating their potential for clinical application.

Tabela 1. Quantitative performance of the models.

Model	F1-score (Macro)	F1-score (Micro)	DP
Gemini 1.5 Flash	0,852	0,853	0,122
GPT-4o	0,797	0,798	0,128
LLaMA 3.3	0,806	0,809	0,105
Mean	0,818	0,820	0,118

3.3.1. Statistical analysis (ANOVA)

To assess whether differences among models were statistically significant, a one-way ANOVA was applied to individual F1-score values ($n = 180$). The test indicated a significant difference between means ($F(2, 177) = 3.75; p = 0.025$).

Gemini 1.5 Flash obtained the highest mean (0.85 ± 0.12), followed by LLaMA 3.3 (0.81 ± 0.11) and GPT-4o (0.80 ± 0.13). Post-hoc comparisons with Bonferroni correction ($\alpha = 0.0167$) revealed a significant difference only between Gemini 1.5 Flash and GPT-4o ($\Delta = +0.06; p = 0.017$) (Table 2).

Tabela 2. Summary of ANOVA results and post-hoc comparisons (Bonferroni).

Comparison	Mean difference	p-value	Significant
Gemini vs GPT-4o	+0,0554	0,0165	Yes
Gemini vs LLaMA	+0,0460	0,0287	No
GPT-4o vs LLaMA	-0,0094	0,6608	No

In summary, Gemini 1.5 Flash performed significantly better than GPT-4o, while LLaMA 3.3 showed intermediate results without relevant differences. The observed difference, although moderate ($\approx$ 5.5 percentage points in F1-score), confirms the robustness and consistency of Gemini 1.5 Flash for report structuring.

Section-level analysis revealed variations associated with the degree of textual standardization and semantic complexity of each part of the template. Mean F1-scores are shown in Table 3.

Tabela 3. Mean F1-score ($\pm$ SD) by section and language model.

Section	Gemini 1.5 Flash	GPT-4o	LLaMA 3.3
Findings	0,845 $\pm$ 0,157	0,849 $\pm$ 0,155	0,790 $\pm$ 0,166
Root	0,797 $\pm$ 0,298	0,797 $\pm$ 0,290	0,858 $\pm$ 0,264
Technique	0,950 $\pm$ 0,220	0,400 $\pm$ 0,494	0,767 $\pm$ 0,427

In the *findings* section, models performed similarly, with GPT-4o (0.849 ± 0.155) and Gemini 1.5 Flash (0.845 ± 0.157) slightly ahead of LLaMA 3.3 (0.790 ± 0.166). This proximity indicates good capability to restructure clinical findings even in highly variable text.

In the *root* section, which includes introductory elements, performance was stable and comparable across all three models, with means around 0.80 ± 0.28 and a slight advantage for LLaMA 3.3 (0.858 ± 0.264).

The largest discrepancy occurred in the *technique* section, where Gemini 1.5 Flash maintained high performance (0.950 ± 0.220) and LLaMA 3.3 showed intermediate performance (0.767 ± 0.427), while GPT-4o dropped markedly (0.400 ± 0.494). This difference resulted from GPT-4o's literal behavior, reproducing the prompt's few-shot example instead of the original text, indicating higher sensitivity to standardized excerpts.

When *technique* section was excluded from the analysis, the overall means became nearly identical (0.82 ± 0.24; 0.82 ± 0.23; 0.82 ± 0.22), respectively for Gemini, GPT-4o, and LLaMA, confirming that the initial discrepancy was associated with the nature of the *technique* section rather than true performance differences among models. Overall, fields with predictable structure and standardized terminology, such as `diagnostic_impression`, `indication`, and `findings.devices_materials`, showed the highest scores (≥ 0.97). In contrast, free-text or contextual fields, such as `incidental_findings` (0.211), `lungs_pleura` (0.554), and `note` (0.651), showed lower performance.

The `technique.description` subfield showed satisfactory performance with Gemini 1.5 Flash (0.949), but significant instability with GPT-4o (0.407), attributed to literal reproduction of a few-shot learning example. The overall mean was 0.806, with Gemini 1.5 Flash achieving the best aggregated result (0.842).

Fields strongly dependent on clinical context or non-standardized language remained the main challenges, reinforcing the role of quantitative analysis in identifying error patterns and optimizing prompts, thus contributing to greater clinical reliability. Text similarity metrics were used only in isolated cases and were not the main focus of this analysis.

3.4. Execution time and processing costs

For the local experiments, LLaMA 3.3 was executed on a workstation with an Intel Core i7 CPU, NVIDIA RTX 3080 GPU, 32 GB of RAM, and a 1 TB NVMe SSD. GPT-4o and Gemini 1.5 Flash were executed via cloud APIs.

Mean processing time per report varied across models, reflecting differences in architecture and execution environment. Locally run LLaMA 3.3 had the lowest mean time (1.49 ± 0.35 s), followed by GPT-4o (1.87 ± 0.51 s) and Gemini 1.5 Flash (2.51 ± 0.73 s). All models completed structuring of the 60 reports without execution failures.

From a cost perspective, processing was performed under a single monthly ChatGPT-4o subscription, with no variable per-report charge. Therefore, costs were treated only as a fixed operational expense and not compared across models. Overall, execution time was low in all cases, reinforcing the practical feasibility of automated report structuring at clinical scale.

3.5. Discussion

In our experiments, GPT-4o and Gemini 1.5 Flash showed similar performance, generating structured reports that were clear, high-fidelity, and linguistically consistent. Both

models correctly interpreted the prompts and filled the template despite variability in the original texts. The main limitations were observed in longer or more subjective fields that require greater semantic precision, a challenge also described in prior studies on radiology report structuring [Takita et al. 2025].

LLaMA 3.3 and Gemini 1.5 Flash tended to leave fields empty without justification, contrary to prompt instructions. GPT-4o showed occasional failures in the *technique* field, literally reproducing few-shot learning examples, reflecting its sensitivity to internal instruction coherence. In more objective and standardized fields, all models maintained fidelity and completeness, with GPT-4o and Gemini 1.5 Flash demonstrating better textual fluency and organization. Simple reports were more accurate, whereas comparative reports required greater interpretive capacity, as observed in prior work [Woźnicki et al. 2024, Shah 2024, Dorfner et al. 2024, Bosbach et al. 2024].

These results are aligned with recent literature: Dorfner et al. [Dorfner et al. 2024] reported strong LLM performance in classification tasks but reduced effectiveness on longer descriptions; Takita et al. [Takita et al. 2025] highlighted the usefulness of classical metrics (precision, recall, and F1-score) while noting that they are insufficient to assess clarity and naturalness; and Woznicki et al. [Woźnicki et al. 2024] suggested complementary measures, such as the Matthews Correlation Coefficient, for imbalanced scenarios.

The most recurrent failures included leaving fields empty, incorrect segmentation, and preservation of spelling errors-findings also described by McFarland et al. [McFarland et al. 2021] and Woznicki et al. [Woźnicki et al. 2024]. No model consistently corrected spelling, which may affect standardization and readability, reinforcing the need for prompt adjustments, template refinement, and automated post-processing validation [Pesapane et al. 2023, McFarland et al. 2021, Russe et al. 2024].

The study has limitations inherent to its exploratory nature, with a qualitative sample of 180 reports (a fraction of the 1,102 available) and data from a single hospital, which restricts generalizability. Models were not supervised fine-tuned in Portuguese, prompts were manually defined, and tests did not include integration with real clinical systems (PACS/RIS) [Woźnicki et al. 2024, Tam et al. 2024].

Even so, findings highlight the potential of LLMs for automated structuring of radiology reports in Portuguese, establishing a foundation for large-scale prospective studies. The combination of objective metrics and qualitative evaluation by specialists supports the feasibility of safely integrating artificial intelligence into radiology practice, considering the linguistic and editorial diversity observed in the Brazilian context.

4. Conclusion

This paper evaluated, using real-world Portuguese chest CT reports, the ability of different LLMs to convert free-text narratives into structured JSON through an adaptable template, emphasizing fidelity, clarity, and structural consistency. The results indicate that Portuguese report structuring is feasible and can yield high-quality, machine-readable outputs without fine-tuning. We synthesize the quantitative and qualitative evidence by answering the three research questions.

RQ1: LLMs can structure Portuguese radiology reports into JSON with high fidelity and consistency, supported by strong quantitative performance (best results with Gemini 1.5

Flash: F1 macro/micro = 0.852/0.853) and positive qualitative assessment by radiologists.

RQ2: A practical trade-off emerged. Proprietary models provided the best overall quality, with Gemini leading quantitatively and GPT-4o receiving the highest radiologist preference (30.08%). In contrast, the locally deployable LLaMA 3.3 achieved competitive metrics but lower clinical preference, suggesting additional engineering is needed to match proprietary outputs.

RQ3: Field-level analysis showed that standardized sections reached very high accuracy (F1≥ 0.97), whereas highly variable narrative fields concentrated errors (e.g., *achados_incidentais* and *pulmoes_pleura*). Prompting and few-shot strategies improved structure and reduced some failures, but variability-driven errors remain key targets for refinement.

Overall, the study provides practical evidence for adopting LLMs to structure Portuguese radiology reports and motivates prospective evaluation with stronger validation and field-specific optimization.

Referências

Adams, L. C., Truhn, D., Busch, F., Kader, A., Niehues, S. M., Makowski, M. R., et al. (2023). Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study. *Radiology*, 307(4):e230725.

Bhayana, R. (2024). Chatbots and large language models in radiology: A practical primer for clinical and research applications. *Radiology*, 310(1):e232756.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., and Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6:149.

Bosbach, W. A., Senge, J. F., Nemeth, B., Omar, S. H., Mitrakovic, M., Beisbart, C., et al. (2024). 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*, 53(1):102–110.

Dorfner, F. J., Jürgensen, L., Donle, L., Mohamad, F. A., Bodenmann, T. R., Cleveland, M. C., et al. (2024). Comparing commercial and open-source large language models for labeling chest radiograph reports. *Radiology*, 313(1):e241139.

Elvas, L. B., Almeida, A., and Ferreira, J. C. (2025). Natural language processing in medical text processing: A scoping literature review. *International Journal of Medical Informatics*, 204:106049.

Fink, M. A., Bischoff, A., Fink, C. A., Moll, M., Kroschke, J., Dulz, L., et al. (2023). Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology*, 308(3):e231362.

Goldberg-Stein, S. and Chernyak, V. (2019). Adding value in radiology reporting. *Journal of the American College of Radiology*, 16(9):1292–1298. Pt B.

McFarland, J. A., Elkassem, A. M. A., Casals, L., Smith, G. D., Smith, A. D., and Gunn, A. J. (2021). Objective comparison of errors and report length between structured and freeform abdominopelvic computed tomography reports. *Abdominal Radiology*, 46(1):387–393.

Mozayan, A., Fabbri, A. R., Maneevese, M., Tocino, I., and Chheang, S. (2021). Practical guide to natural language processing for radiology. *RadioGraphics*, 41(5):1446–1453.

Nobel, J. M., Geel, K. V., and Robben, S. G. F. (2022). Structured reporting in radiology: a systematic review to explore its potential. *European Radiology*, 32(4):2837–2854.

Pesapane, F., Tantrige, P., Marco, P. D., Carriero, S., Zugni, F., Nicosia, L., et al. (2023). Advancements in standardizing radiological reports: A comprehensive review. *Medicina*, 59(9):1679.

Russe, M. F., Reisert, M., Bamberg, F., and Rau, A. (2024). Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der Bildgebenden Verfahren*, pages a–2264–5631.

Shah, S. V. (2024). Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953.

Spandorfer, A., Branch, C., Sharma, P., Sahbaee, P., Schoepf, U. J., Ravenel, J. G., et al. (2019). Deep learning to convert unstructured CT pulmonary angiography reports into structured reports. *European Radiology Experimental*, 3(1).

Takita, H., Walston, S. L., Mitsuyama, Y., Watanabe, K., Ishimaru, S., and Ueda, D. (2025). Comparative performance of large language models in structuring head CT radiology reports: multi-institutional validation study in japan. *Japanese Journal of Radiology*.

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., et al. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1):258.

Woźnicki, P., Laqua, C., Fiku, I., Hekalo, A., Truhn, D., Engelhardt, S., et al. (2024). Automatic structuring of radiology reports with on-premise open-source large language models. *European Radiology*, 35(4):2018–2029.