

Extraction of clinical findings from mammography and breast ultrasound reports: a comparison between specialists and Artificial Intelligence

Lorenzo Farias ^{1,*}, Hanna Reckziegel ¹, Daniela Duarte da Silva Bagatini ^{1,2}, Daniel Schulz ¹, Gabriela de Andrade Monteiro ³, Leticia Zanatta ⁴, Ana Laura Brill Thum ⁵, Priscila Schmidt Lora ⁵, Débora Oliveira da Silva ³, Ana Paula Wernz da Cunha Müller ⁶, and Cristiane Drebes Pedron ⁷

¹Department of Engineering, Architecture and Computing, Universidade de Santa Cruz do Sul (UNISC), Santa Cruz do Sul, RS, Brazil

²Graduate Program in Industrial Systems and Processes (PPGSPI), Universidade de Santa Cruz do Sul (UNISC), Santa Cruz do Sul, RS, Brazil

³Graduate Program in Production and Systems Engineering (PPGEPS), Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, RS, Brazil

⁴Faculty of Medicine (FaMed), Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brazil

⁵School of Health, Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, RS, Brazil

⁶Clínica de Mastologia Dra. Ana Paula Muller Ltda., Porto Alegre, RS, Brazil

⁷Graduate Program in Administration (PPGA) and Graduate Program in Project Management (PPGP), Universidade Nove de Julho (UNINOVE), São Paulo, SP, Brazil

*Corresponding author: fariaslorenzo@mx2.unisc.br

Abstract

Breast cancer is the leading cause of cancer-related death among women in Brazil, and the time between the request and the release of mammography reports directly influences adherence to screening, making the agility in processing these reports a critical factor for early diagnosis. In this context, this study compares the performance of a Large Language Model (LLM) with manual extraction performed by a team of health researchers in identifying clinical findings from mammography and breast ultrasound reports written in Brazilian Portuguese. Named Entity Recognition (NER) was applied through Prompt Engineering using a *few-shot* strategy, employing the Gemini 2.5 Flash model, selected from preliminary exploratory tests with four candidate models. The Gemini 2.5 Flash model demonstrated the best performance, achieving a Macro F1 of 0.91 and a Micro F1 of 0.98. The subjective validation, in which 29 exams — of different formats — were evaluated by health researchers using a Likert scale, yielded an agreement index of 93.1%. The model outperformed human extraction in overall Macro F1 (0.91 vs. 0.72), as in four reports the model correctly identified information that had been omitted or incorrectly recorded during manual extraction, demonstrating its potential as a complementary verification tool alongside specialists. The results confirm the hypothesis that LLMs, when instructed through Prompt Engineering, can achieve performance comparable to or superior to manual extraction by health professionals.

Keywords: Mammography; Prompt Engineering; Named Entity Recognition; Natural Language Processing; Large Language Models

1 Introduction

Breast cancer is the leading cause of cancer-related death among women in Brazil, accounting for 16.5% of oncological deaths in the country [1]. The estimate for the 2023–2025 triennium is 73,610 new cases annually, with an adjusted incidence rate of 41.89 cases per 100,000 women [2].

In Rio Grande do Sul, 5,038 cases were recorded in 2023, a 35.4% increase over the estimated value, representing the third highest number in the country [3, 4].

In the breast cancer care pathway, the time between the request and the release of reports directly influences adherence to screening. In 2023, only 48.8% of screening mammography reports were released within 30 days, while approximately 36% took more than 60 days [1]. Furthermore, the lack of interoperability between health systems compromises continuity of care and may negatively affect clinical outcomes [5–7]. This scenario highlights the need for process improvements to expedite feedback to patients, thereby strengthening confidence in screening programs.

The use of Artificial Intelligence (AI) can bring numerous benefits, such as resource optimization, faster diagnoses, and error reduction [8–11]. One approach to mitigating the latency between examination and patient comprehension of results consists of automating the extraction of relevant clinical findings from diagnostic reports, using the concept of Named Entity Recognition (NER). In the healthcare context, NER automatically extracts information such as diagnoses, medications, or dates from reports and medical records, supporting data structuring.

Considering this context, the present study is based on the hypothesis that LLMs, when correctly instructed through Prompt Engineering techniques, can achieve performance comparable to manual extraction of clinical findings performed by specialists. The selection of the LLM model used was conducted through preliminary exploratory tests, detailed in the Methods section.

The main objective of this work is to compare the performance of an LLM model with manual extraction performed by health researchers in identifying pre-selected clinical findings in mammography and breast ultrasound reports. To this end, the following specific objectives were defined: (1) mapping, together with health researchers and a specialist, the relevant clinical finding present in the reports; (2) developing and refining a structured *prompt* based on identified patterns; (3) quantitatively evaluating the automated extraction through classification metrics (Accuracy, Precision, Recall, and F1-Score); and (4) qualitatively validating the results with health professionals through the Likert scale.

This article contributes to the field of digital health by providing empirical evidence that LLMs, when adequately instructed, can serve as a clinical support tool for structuring data from non-standardized reports written in Brazilian Portuguese.

2 Related Work

AI-based solutions have been proposed for the automated extraction of clinical data, notably BioBERT [12], which employs the BERT (*Bidirectional Encoder Representations from Transformers*) architecture, and other NLP techniques. However, these approaches require extensive training datasets with local clinical terminology, which constitutes a considerable challenge for different linguistic and cultural contexts.

Currently, LLMs demonstrate excellent results across multiple languages and greater adaptability with less need for specific training [13–17]. Godoy *et al.* [14] evaluated three OpenAI models (GPT-4o, GPT-4-turbo, and GPT-3.5-turbo) for automatic laterality detection in 377 mammography reports in Spanish, using *zero-shot*, *one-shot*, and *few-shot* strategies. GPT-4o with the *few-shot* approach achieved an F1 of 0.77, significantly outperforming GPT-3.5-turbo. The analysis identified three types of errors: positioning (Type 1), format (Type 2), and severe non-correctable errors (Type 3).

Chen *et al.* [17] investigated three Chinese LLMs with fine-tuning via LoRA for structuring 550 breast ultrasound reports. Baichuan2-7b achieved 97.63% precision for structure extraction and 96.22% for attribute-value matching. The study demonstrated that dividing reports into logical sections significantly improved performance compared to holistic structuring, with 400 training samples being the optimal volume.

Akali *et al.* [13] applied a many-shot prompt engineering strategy using Google’s Gem-

ini 1.5 Pro to automatically extract five clinical entities — anatomy, impression, and observation presence, absence, and uncertainty — from 85 Turkish mammography reports, a language with limited available NLP resources. The method incorporated 165 annotated examples within a 26,000-token prompt, yielding a macro-averaged F1 of 0.99 under relaxed matching and 0.84 under exact matching. The study demonstrated that many-shot prompting with a large-context-window LLM consistently outperforms zero-shot and few-shot strategies for non-English medical report annotation, reinforcing the adaptability of prompt engineering to languages beyond English without requiring task-specific model training.

3 Methods and Procedures

This study adopts an iterative experimental design, combining quantitative performance evaluation with qualitative expert validation. Clinical variables were defined in collaboration with health researchers and a mastologist. The *prompt* development followed an incremental process — evaluating multiple LLM models, input formats, and prompting strategies — to identify the best-performing configuration. The final system was evaluated using classification metrics (Accuracy, Precision, Recall, and F1-Score) against a manually adjudicated ground truth, complemented by qualitative validation via Likert scale. The project follows the *Design Science Research* (DSR) approach [18, 19], structured in five stages [20], described in the following subsections.

3.1 Procedures

Stage 1 – Problem Awareness: a literature review was conducted in PubMed, Scopus, Web of Science, and SciELO databases, which revealed gaps in the automated extraction of clinical findings from mammography reports in Brazilian Portuguese and the relevance of report release delays for screening adherence. Based on this review, the breast cancer care pathway at the screening stage was selected as the focus of this study.

Stage 2 – Suggestion: mapping, together with health professionals (mastologist and medical students), of clinical variables (cysts, nodules, calcifications, microcalcifications, and BI-RADS) and AI technologies (RAG, Prompt Engineering, Fine-Tuning, NLP, and LLMs).

Stage 3 – Development: exploratory tests evaluating (1) PDF vs. TXT formats [21]; (2) *Zero-shot*, *Few-shot*, and *Multi-shot* strategies; and (3) GPT-4o, Claude 3.5 Sonnet, DeepSeek R1, and Gemini 2.5 Flash models. RAG was compared, but Prompt Engineering demonstrated sufficient efficacy with lower complexity. A Python 3.13 application was developed for extraction in structured JSON.

Stage 4 – Evaluation: the evaluation was conducted by comparing the results obtained by the AI with those produced by the researchers, using quantitative metrics (Accuracy, Sensitivity, Specificity, and F1-Score) and a qualitative measure represented by the Likert Scale. The F1-Score was calculated for both human evaluators and the AI, while the Likert Scale was used to measure the degree of agreement of the health professional with the extractions performed by the model.

Stage 5 – Conclusion and Communication: results reported in this article.

3.2 Study Sample

The sample consisted of 33 examinations: 29 for testing and comparing *prompts*, and 4 selected as representative examples to compose the final *prompt* in the chosen *few-shot* approach. The examinations were obtained from the mastology clinic partnering with the project, covering reports from different laboratories and radiology services in Rio Grande do Sul, with variations in terminology and writing style, which increases the diversity of the sample.

3.3 Data Mapping

Figure 1 illustrates the clinical data mapping stage from examination reports. The collected reports were subjected to analysis by health researchers, who performed the extraction of relevant findings to construct a structured reference informational model in a spreadsheet. This process underwent agreement evaluation between the specialist and health researchers to ensure consistency and reliability of the extracted data.

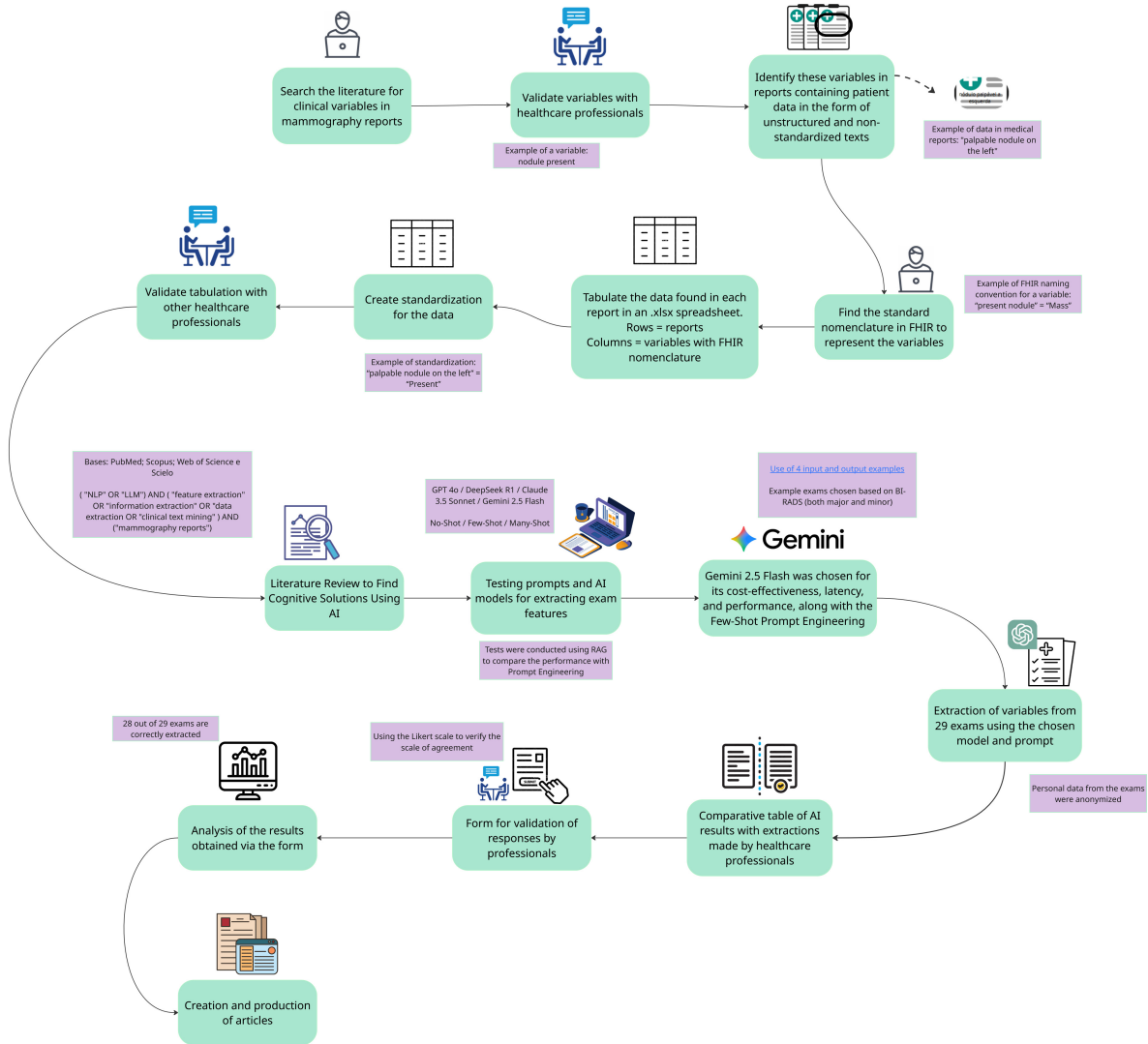


Figure 1: Mapping process flowchart. Illustrates the stages from bibliographic search for clinical variables, validation with health professionals, identification in reports, FHIR standardization, tabulation in spreadsheet, and final validation.

3.3.1 AI Model Selection

Model selection was conducted through exploratory tests with four candidates evaluated for qualitative performance, API access cost, response latency, and score on the LMArena benchmark [22], a platform that compares models for general language processing capability. Table 1 summarizes the characteristics of the evaluated models.

Table 1: Characteristics of models evaluated in exploratory tests. Prices in USD per million tokens. *Benchmark*: LMArena score [22].

Model	Benchmark	Latency (s)	Input P.	Output P.
GPT-4o (OpenAI)	1408	0.29	\$2.50	\$10.00
Claude 3.5 Sonnet (Anthropic)	1283	0.95	\$3.00	\$15.00
DeepSeek R1	1359	3.40	\$1.35	\$5.40
Gemini 2.5 Flash (Google)	1394	8.15	\$0.15	\$3.50

Gemini 2.5 Flash was selected for presenting the second highest benchmark score (1,394 points), significantly lower input cost (\$0.15/M tokens), API access availability through Google during the research period, and best qualitative performance in exploratory tests. Although its average latency (8.15 s) is higher than the other models, this factor does not constitute a relevant restriction in the context of batch processing of clinical reports.

3.3.2 Parameter Definition and *Prompt* Testing

Three prompt engineering approaches were evaluated: (1) **Zero-shot** – only the task instruction, with no prior examples; (2) **Few-shot** – from 1 to 5 examples of analyzed reports; and (3) **Multi-shot** – 15 representative examples. The extractions generated by the AI were tabulated and compared to the gold standard established by the multidisciplinary team of health researchers. After performance comparison, the *few-shot* strategy with 4 examples was selected. As observed in lightweight or distilled models (such as Gemini 2.5 Flash or smaller local models), while *zero-shot* suffered from structural inconsistency in the JSON output, *multi-shot* led to attention dilution [23]. Although these models possess large context windows, providing an excessive number of examples exacerbates the “lost in the middle” phenomenon [24] and overwhelms their attention mechanisms. This “over-prompting” degrades extraction quality and increases the hallucination rate [23], as the model’s attention is spread too thin across the examples, causing it to confuse or merge clinical details from the examples with the actual report being analyzed. Therefore, *few-shot* presented the best technical balance.

Regarding data input, PDF files can induce hallucinations [21] in AI models due to their complex structure (metadata, tables, and images that may be interpreted inconsistently). To mitigate this issue, the TXT format was adopted for report input, preserving only the essential textual content.

Figure 2 illustrates the *prompt* testing workflow up to the use of the final prototype version.

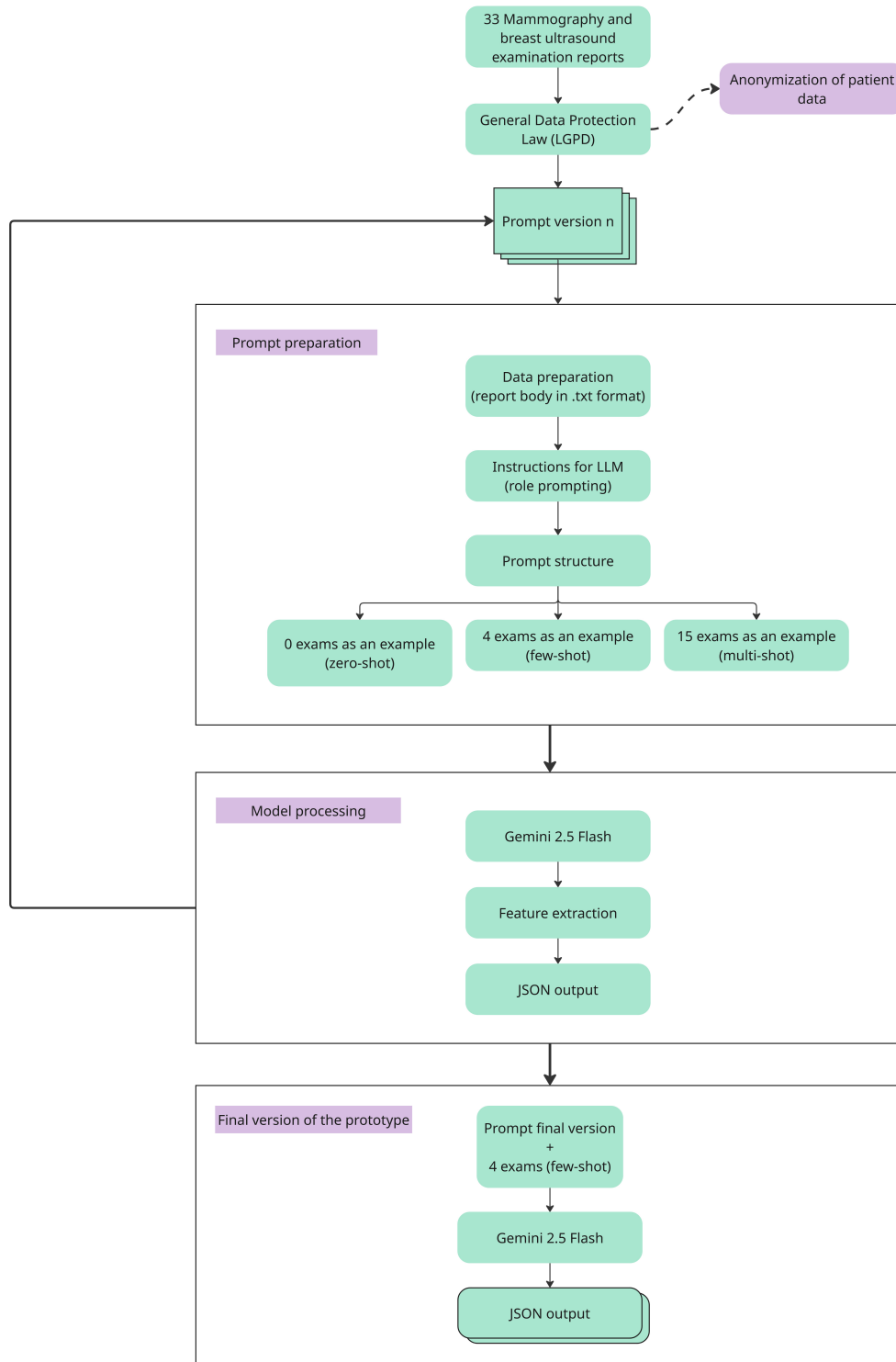


Figure 2: Technical methodology flowchart. Shows the stages from report anonymization to *prompt* preparation, model processing, JSON output, and use in the final prototype.

3.3.3 Comparative Analysis and Statistical Evaluation

For qualitative analysis, the extractions performed by the AI were organized in a spreadsheet, where each row represents an examination and the columns represent clinical variables. A specialist and a health researcher received the original reports along with the extraction spreadsheet for cross-validation. The Likert scale was used in response to the guiding question: “*Regarding this report, are the presented answers correct?*”

For quantitative analysis, the F1-Score (harmonic mean of Precision and Recall) was adopted as the central metric, widely used in NLP tasks [12, 13]. The metrics were calculated from: **TP** (AI correctly identifies a present finding); **TN** (AI correctly identifies an absence); **FP** (AI identifies something nonexistent – hallucination); and **FN** (AI omits a present finding):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

For establishing the Ground Truth, the manual extraction by health researchers was used, subjected to an adjudication process in cases of disagreement between manual extraction and AI extraction. In such cases, the multidisciplinary team, composed of the same health researchers involved in the initial extraction, returned to the original report for joint review, determining by consensus which result was compatible with the explicit content of the document. The ground truth resulting from this process was adopted as the reference for calculating all quantitative metrics.

4 Results

The performance evaluation of the Gemini 2.5 Flash model was conducted on two fronts: objective quantitative analysis (F1-Score and confusion matrices) and subjective qualitative analysis (agreement by researchers and specialist).

4.1 Quantitative Performance and Adjudication

In the automated evaluation of 29 test reports, the model demonstrated high efficacy in the identification and extraction of clinical entities. During validation, it was observed that the initial Gold Standard (manual extraction) contained inaccuracies in four examinations (IDs 1, 10, 16, and 70). After adjudication, the AI results proved to be correct.

The system evaluated 116 binary instances (presence/absence of Cysts, Nodules, Calcifications, and Microcalcifications in 29 reports), recording 39 True Positives (TP), 75 True Negatives (TN), 1 False Positive (FP), and 1 False Negative (FN). Table 2 presents performance by clinical entity, with Macro F1 of 0.91 and Micro F1 of 0.98.

Table 2: Performance per clinical entity of the Gemini 2.5 Flash model after gold standard adjudication.

Clinical Entity	Accuracy	Sensitiv.	Specific.	Precision	F1
Cysts	1.00	1.00	1.00	1.00	1.00
Nodules	1.00	1.00	1.00	1.00	1.00
Calcifications	0.97	0.94	1.00	1.00	0.97
Microcalcifications	0.97	1.00	0.96	0.50	0.66
Macro F1	0.98	0.98	0.99	0.87	0.91
Micro F1^a	—	0.98	—	0.98	0.98

^aMicro F1: TP = 39, FP = 1, FN = 1 (Precision = Recall = 0.975).

As the BI-RADS category constitutes the only multiclass variable (categories 1 to 6), its results are presented separately in Figures 3 and 4.

Figure 3 presents the global confusion matrix for binary entities. Of the 116 evaluated instances, the model recorded 39 True Positives (TP) and 75 True Negatives (TN), with only 1 False Positive (FP) and 1 False Negative (FN), evidencing high discriminative capacity for both the presence and absence of clinical entities.

Figure 4 displays the confusion matrix for BI-RADS categories 1 to 6. All 29 examinations were correctly classified on the main diagonal — BI-RADS 2 was the most frequent category ($n = 15$), followed by BI-RADS 4 ($n = 5$), BI-RADS 1 ($n = 4$), BI-RADS 5 ($n = 3$), and BI-RADS 6 ($n = 1$) —, with no confusion between classes, resulting in Accuracy of 1.00. This result is clinically relevant, as BI-RADS classification errors can imply inadequate clinical management, from delayed necessary biopsies to unnecessary interventions.

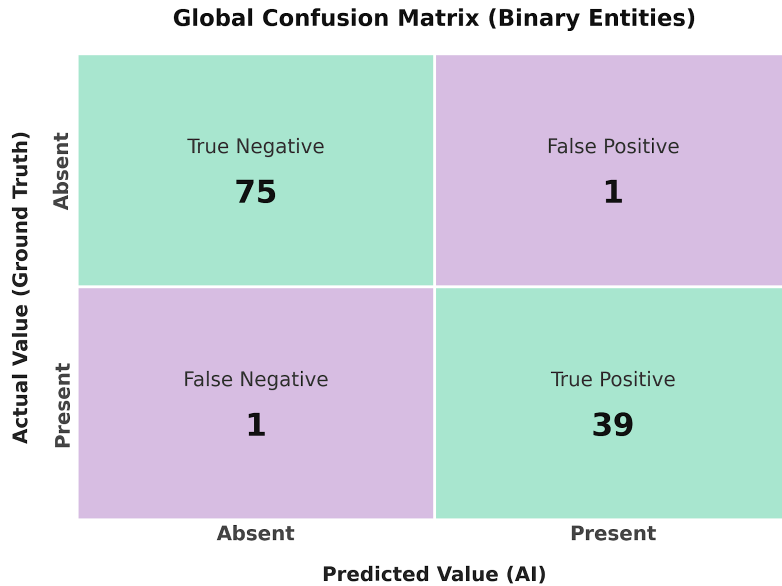


Figure 3: Global Confusion Matrix (Binary Entities). The model recorded 75 TN and 39 TP, with only 1 FP and 1 FN across 116 evaluated instances.

Confusion Matrix: BI-RADS Categories							
Actual Category (Ground Truth)	BI-RADS 1	4	0	0	0	0	0
	BI-RADS 2	0	15	0	0	0	0
	BI-RADS 3	0	0	1	0	0	0
	BI-RADS 4	0	0	0	5	0	0
	BI-RADS 5	0	0	0	0	3	0
	BI-RADS 6	0	0	0	0	0	1
		BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5	BI-RADS 6
Predicted Category (AI)							

Figure 4: Confusion Matrix for the BI-RADS Category. All 29 examinations were correctly classified in categories 1 to 6 (Accuracy = 1.00), with no overlap between classes.

4.2 Comparison Between AI and Human

Table 3 presents the comparative performance between the human (pre-adjudication manual extraction: TP = 37, TN = 75, FP = 1, FN = 3) and the Gemini 2.5 Flash model, with the adjudicated ground truth adopted as the gold standard in both cases.

Table 3: Comparison of F1-Score and Sensitivity per clinical entity between human and Gemini 2.5 Flash (adjudicated ground truth as gold standard).

Clinical Entity	F1-Score		Sensitivity	
	Human	AI	Human	AI
Cysts	1.00	1.00	1.00	1.00
Nodules	0.90	1.00	0.90	1.00
Calcifications	0.97	0.97	0.94	0.94
Microcalcifications	0.00 ^a	0.66	0.00	1.00
Macro F1	0.72	0.91	0.71	0.98
Micro F1	0.95	0.98	0.93	0.98

^aPrecision undefined (TP = 0, FP = 0); F1 = 0. Human Micro F1: TP = 37, FP = 1, FN = 3 (Precision = 0.97; Recall = 0.93).

The most expressive difference was concentrated in the Microcalcifications category. The only real case of this entity in the test set was detected by the AI (sensitivity = 1.00), while the human did not flag it (sensitivity = 0.00). The human accumulated three omissions (FN = 3) against only one by the AI (FN = 1), indicating greater human susceptibility to omission errors in systematic report review. The AI Macro F1 (0.91) outperformed that of the human (0.72) by 0.19 points. The Micro F1, by weighting by instance volume, places the performances closer together (0.98 vs. 0.95).

Qualitative analysis of errors and correct extractions revealed that the AI acted as a verification tool (*double-check*), mitigating human attention errors. The main observations were:

- **Correction of Human False Positives (Specificity):** In Exam 1, manual extraction incorrectly indicated the presence of a nodule. The AI correctly identified its absence (specificity = 1.00).
- **Correction of Human Omissions (Sensitivity):** In Exams with identifier numbers 10, 16, and 70, the model detected calcifications and nodules omitted in the manual tabulation. Additionally, the only case of microcalcification in the test set was identified by the AI and not by the human.
- **Taxonomic Error:** In Exam 4, the model classified a calcification as a microcalcification (dimensional criterion < 0.5 cm vs. the human morphological criterion), generating 1 FP and 1 FN. This behavior is correctable through prompt adjustment.

4.3 Subjective Validation (Agreement)

To measure qualitative performance, the Global Acceptance Score was adopted, converting Likert scale responses to numerical values ('Strongly Agree' = 100; 'Strongly Disagree' = 0). The final index obtained was 93.1% (Figure 5), based on 58 assessments conducted by two evaluators, with each one responsible for 29 exams.

Of the 58 evaluations, 75.9% ($n = 44$) corresponded to "Strongly Agree" and 22.4% ($n = 13$) to "Partially Agree". Only one evaluation (1.7%) indicated partial disagreement, and no total disagreement was recorded.

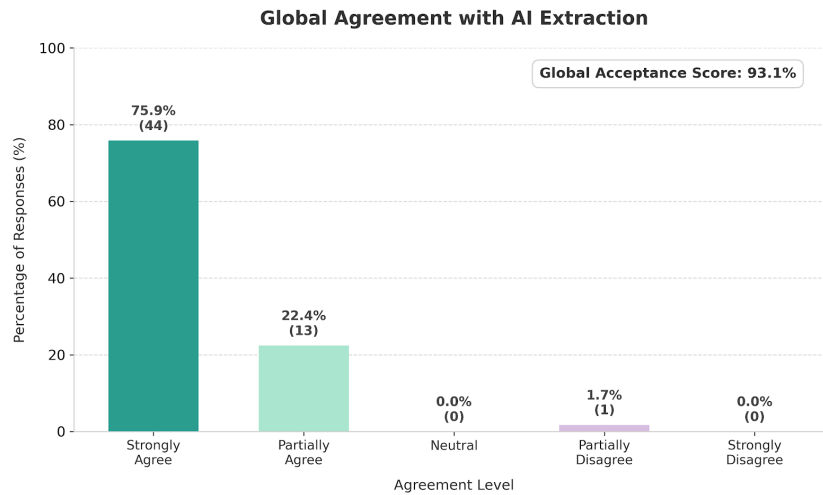


Figure 5: Level of professional agreement with AI. The Global Acceptance Index calculated over 58 evaluations (Likert scale, two professionals) was 93.1%.

The results demonstrated that the AI presented significant efficacy in automating the extraction process, with greater completeness of results, lower probability of omissions, and superior speed. The professionals demonstrated advantages in interpretive flexibility for atypical cases, such as Exam 4.

Figure 6 presents the levels of agreement between professionals by examination. The color scale reflects the degree of consensus: dark green indicates full agreement; light green denotes subtle divergence (between 'strongly agree' and 'partially agree'); and lilac signals disagreement. Only in Exam 4 was there interpretive divergence regarding the classification of the punctiform calcification.

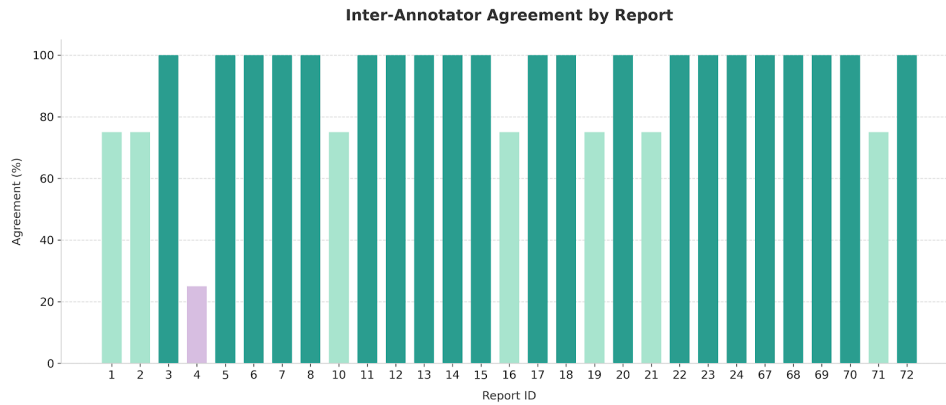


Figure 6: Levels of agreement between professionals by examination. Dark green = full agreement (100%); light green = subtle divergence; lilac = disagreement. Only Exam 4 showed substantive divergence regarding the punctiform calcification.

5 Discussion

5.1 Error Analysis

There were two residual errors mitigable through prompt adjustment. Figure 7 details the inconsistency in Exam 4, where the model classified a punctiform calcification as a microcalcification (strict size criterion: < 0.5 cm), while the professionals adopted morphological and clinical criteria. This disagreement is correctable through prompt adjustments.

Gemini 2.5 Flash NS		CISTO		NÓDULO		CALCIFICAÇÃO		MICROCALCIFICAÇÃO		BIRADS	OUTRAS CITAÇÕES
Exame 4		PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO		
Especialista humano		presente	(i) junção dos quadrantes internos à direita (ii) isolados bilaterais (iii) menores que 1,0cm de diâmetro.	presente	(i) nódulo oval, paralelo, contendo cisto em seu interior (ii) 1,3 cm (iii) localizado em junção dos quadrantes internos à direita (estável).	presente (calcificações puntiformes, tênues, agrupadas)	em uma área de aproximadamente 0,5 cm, localizadas em região central a esquerda	ausente		4	Na conclusão: (i) as calcificações de mama esquerda envolvem suspeita. (ii) Ectasia ductal retroareolar bilateral. (iii) axilas livres.
IA		Sim	Cistos no interior de nódulo complexo (1,3cm) em junção dos quadrantes internos à direita; Cistos isolados bilaterais menores que 1,0cm de diâmetro.	Sim	Nódulo oval, paralelo, complexo (contendo cistos), medindo 1,3cm, localizado em junção dos quadrantes internos à direita (estável).	Não		Sim	Calcificações puntiformes, tênues, agrupadas em uma área de aproximadamente 0,5cm, localizadas em região central à esquerda.	4	Mamas heterogeneamente densas; As calcificações de mama esquerda sofreram aumento numérico significativo em relação ao exame anterior de 03/2022 e envolvem suspeita; Ectasia ductal retroareolar bilateral; Axilas livres.

Figure 7: Exam 4: comparison between AI extraction (Gemini 2.5 Flash) and human. The AI applied a dimensional criterion (< 0.5 cm) to classify the calcification as a microcalcification, while the human adopted morphological and clinical criteria. Disagreement correctable through prompt adjustment.

Exam 15 (Figure 8) was prioritized in tests for inducing hallucinations in PDF format. In .txt format, full agreement was achieved across all clinical variables. This result is consistent with the literature indicating that the quality of input preprocessing directly influences the fidelity of LLM responses [21], and with practical studies demonstrating that structural elements of PDFs — such as tables, metadata, and complex layouts — represent significant challenges for text extraction intended for AI pipelines, favoring the use of plain text formats.

Gemini 2.5 Flash NS	CISTO		NÓDULO		CALCIFICAÇÃO		MICROCALCIFICAÇÃO		BIRADS	OUTRAS CITAÇÕES
Exame 15	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO		
Especialista humano	presente	cisto oleoso na mama esquerda	presente		presente		ausente			Alterações arquitetais, relacionadas à mastoplastia. Implante bilateral, sem sinais de roturas extracapsulares. Status pós cirurgia conservadora da mama esquerda.
IA	Sim	Cisto oleoso na mama esquerda.	Sim	Nódulo denso de contornos espiculados projetado no QSE da mama esquerda, associado a retração cutânea, com correspondência ao ultrassom, maior em relação ao exame de 01/2024. Prosseguir com core biopsy.	Sim	Calcificações esparsas.	Não			Indicação clínica: 69 anos. Rotina. Antecedente de neoplasia mamária. Status pós cirurgia conservadora da mama esquerda. Parênquima mamário heterogeneamente denso, o que reduz a sensibilidade da mamografia. Alterações arquitetais, relacionadas à mastoplastia. Implante bilateral, sem sinais de roturas extracapsulares. Linfonodo axilar, de aspecto reacional.

Figure 8: Exam 15: comparison between AI and human extraction. Report selected for its tendency to induce hallucinations in PDF format; full agreement achieved across all variables with .txt input.

Exams 6 and 14 (Figures 9 and 10) are representative of the expected model behavior under normal conditions. In both cases, the extraction performed by the AI was fully consistent with the gold standard, correctly identifying the presence and absence of each clinical entity and the corresponding BI-RADS classification. These results corroborate the findings of Akcali *et al.* [13], who reported Macro F1 between 0.84 and 0.99 for similar tasks in mammography reports in non-English languages, indicating that the proposed approach is competitive even without model fine-tuning and with a reduced example corpus.

Gemini 2.5 Flash NS	CISTO		NÓDULO		CALCIFICAÇÃO		MICROCALCIFICAÇÃO		BIRADS	OUTRAS CITAÇÕES
Exame 6	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO		
Especialista humano	ausente		presente	medindo 0,4 x 0,7 x 0,2cm	ausente		ausente			(i) Mamas com ecotextura heterogênea (ii) implante mamário bilateralmente
IA	Não		Sim	área heterogênea e irregular, no QSL da mama esquerda, medindo 0,4 x 0,7 x 0,2cm.	Não		Não			Mamas com ecotextura heterogênea. Implantes mamários bilateralmente. Sugerimos, a critério clínico, correlação cito-histológica. Resumo clínico: nódulo palpável à esquerda. Não há evidência de outra alteração significativa nas mamas.

Figure 9: Exam 6: comparison between AI and human extraction. Reference report in tests, with full agreement – absence of cyst, calcification, and microcalcifications, and presence of nodule with description.

Gemini 2.5 Flash NS	CISTO		NÓDULO		CALCIFICAÇÃO		MICROCALCIFICAÇÃO		BIRADS	OUTRAS CITAÇÕES
Exame 14	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO	PRESENTE	LOCALIZAÇÃO E TAMANHO		
Especialista humano	ausente		presente	medindo 1,0 x 0,9 x 0,6 cm, distando 2,4 cm do peitoral	ausente		ausente			Mamas predominantemente lipossustituídas. Linfonodos axilares de aspecto normal.
IA	Não		Sim	Nódulo oval e de bordas angulares (não circunscrito) no terço médio de UQS da mama direita, medindo 1,0 x 0,9 x 0,6 cm, distando 2,4 cm do peitoral.	Não		Não			Indicação clínica: biópsia da mama direita em 13/01/2025 (AP: carcinoma invasivo do tipo não especial). Mamas predominantemente lipossustituídas (A). Linfonodos axilares de aspecto normal. Neoplasia confirmada.

Figure 10: Exam 14: comparison between AI and human extraction, demonstrating full agreement including BI-RADS 6 classification and complete nodule description.

5.2 Limitations

The main limitations involved: (1) a dataset limited to 33 examinations from a single partner mastology clinic, in compliance with the ethical restrictions imposed by the General Personal Data Protection Law (LGPD – Law No. 13.709/2018); (2) specificity of the *prompt*, developed and

calibrated from the linguistic and structural patterns of Mammography and Breast Ultrasound examinations only; and (3) low adherence of researchers to the validation form, which was sent to the research team (3 members) and the partner mastologist, resulting in only two respondents.

6 Conclusion

This study concludes that LLMs, when instructed through Prompt Engineering, demonstrate effective capacity to assist in the extraction and summarization of clinical findings from mammography and breast ultrasound reports. Prompt Engineering proved to be a computationally low-cost solution for structuring semi-structured data in a clinical environment.

The results confirmed the hypothesis that LLMs can achieve performance comparable to or superior to manual extraction by specialists, as evidenced by the Macro F1 of 0.91 and Micro F1 of 0.98, compared to the human Macro F1 of 0.72 and Micro F1 of 0.95. The finding that the AI detected the only case of microcalcification present in the test set, while the human missed it, reinforces the model’s potential as a complementary *double-check* tool, especially for low-prevalence entities. However, this omission should not be interpreted as a limitation exclusive to the professional’s technical competence, but rather as a reflection of the inherent conditions of clinical practice. Human attention is a finite resource susceptible to cognitive fatigue, which tends to intensify over long work shifts, in the face of high patient volumes, or in contexts of institutional pressure. The subjective validation (Likert scale, 93.1% acceptance) corroborated these results.

A notable aspect of this study is the model’s robustness against real-world data variability. Since the sample included reports from various laboratories and radiology services in Rio Grande do Sul, which naturally present different formatting, nomenclature, and writing styles, the high performance achieved demonstrates the flexibility of the solution. This suggests that the developed *prompt* architecture is highly adaptable and has great generalization potential, being applicable for data extraction from virtually any mammography and breast ultrasound report written in Brazilian Portuguese, regardless of the issuing institution.

For subsequent studies, the following are planned: (1) increasing the data corpus for more robust validation; (2) adapting the *prompt* for different examination types and report standards; and (3) including a greater number of health professionals in the validation of AI responses.

7 Ethics Statement

This study and the acquisition of reports comply with the General Personal Data Protection Law (LGPD – Law No. 13.709/2018). The examinations were obtained through a Mastology Clinic, with approval by the Research Ethics Committee (CEP) via the Plataforma Brasil platform, under CAAE No. 88036825.9.0000.5343. Included were digital reports of mammography and breast ultrasound examinations that met the following criteria: patient authorization through signature of the Informed Consent Form (ICF); and compliance with current data protection regulations. To preserve patient privacy, personal identification data were removed or anonymized prior to AI analysis, in accordance with research ethical principles and the provisions of the LGPD. Reports that were incomplete, in formats unsuitable for processing, or whose patients did not consent to participation in the research were excluded.

Data Availability Statement

The dataset used in this study consists of mammography and breast ultrasound reports obtained from a partner mastology clinic under the approval of the Research Ethics Committee (CEP), CAAE No. 88036825.9.0000.5343, in compliance with the General Personal Data Protection

Law (LGPD – Law No. 13.709/2018). Due to patient privacy constraints and applicable data protection regulations, the dataset is not publicly available. Requests regarding access to the data may be directed to the corresponding author.

Declaration of Competing Interests

The authors declare that Clínica de Mastologia Dra. Ana Paula Muller Ltda. provided the reports used in this study. Dr. Ana Paula Wernz da Cunha Müller served as an external evaluator in the qualitative validation stage (Likert scale) and assisted in defining the variable search process together with the health researchers, without involvement in the remaining stages of the research. The external collaborators had no influence on the study design, data collection, analysis and interpretation, manuscript writing, or the decision to submit it for publication. The remaining authors declare no competing interests.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

The authors declare that generative AI tools were used in a limited and supervised capacity during the preparation of this manuscript. Specifically, Claude Sonnet 4.6 (Anthropic) was used for grammar correction and linguistic improvement. No portion of the text was generated by AI. All ideas, arguments, and written content are exclusively of human authorship. After each AI-assisted revision, the authors thoroughly reviewed and approved the modified text to ensure its accuracy, coherence, and compliance with the study’s objectives.

Acknowledgements

Funding: This work was developed within the scope of the project “Cognitive solution for identifying findings of examinations in the breast cancer care pathway: screening and pre-diagnosis”, affiliated with the Universidade de Santa Cruz do Sul (UNISC). The research is financially supported by the National Council for Scientific and Technological Development (CNPq) and the Department of Science and Technology (Decit/SECTICS/MS), under CNPq/Decit Call No. 21/2023 – Transdisciplinary Studies in Collective Health.

CRedit Authorship Contribution Statement

Lorenzo Farias: Conceptualization, Methodology, Software, Validation, Writing – Original Draft. **Hanna Reckziegel:** Conceptualization, Methodology, Software, Validation, Writing – Original Draft. **Daniela Duarte da Silva Bagatini:** Conceptualization, Methodology, Supervision, Validation, Writing – Review & Editing. **Daniel Schulz:** Validation. **Gabriela de Andrade Monteiro:** Methodology, Writing – Review & Editing. **Letícia Zanatta:** Validation. **Ana Laura Brill Thum:** Validation. **Priscila Schmidt Lora:** Methodology, Validation. **Débora Oliveira da Silva:** Supervision, Funding Acquisition. **Ana Paula Wernz da Cunha Müller:** Validation, Resources. **Cristiane Drebes Pedron:** Validation.

References

- [1] Instituto Nacional de Câncer (Brasil). Controle do câncer de mama no Brasil: dados e números: 2024. Technical report, INCA, Rio de Janeiro, 2024. URL <https://ninho.inca.gov.br/jspui/bitstream/123456789/17002/1/Controle%20do%20c%20mama%20no%20Brasil%202024.pdf>

- C3%A2ncer%20de%20mamano%20Brasil%20-%20dados%20e%20n%C3%BAmeros%202024.pdf. [accessed 10 December 2025].
- [2] Instituto Nacional de Câncer (Brasil). Estimativa 2023: incidência do câncer no Brasil. Technical report, INCA, Rio de Janeiro, 2022. URL <https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa>. [accessed 10 December 2025].
 - [3] Rio Grande do Sul. Secretaria da Saúde. Observatório do câncer RS, 2025. URL <https://observatoriodocancer.saude.rs.gov.br>. [accessed 8 April 2025].
 - [4] F. M. Vasconcellos et al. Boletim epidemiológico da situação do câncer de mama no estado do Rio Grande do Sul. Technical report, Secretaria de Estado da Saúde/RS, Porto Alegre, 2024. URL <https://admin.estado.rs.gov.br/upload/arquivos/202410/boletim-epidemiologico-cancer-de-mama-no-estado-do-rio-grande-do-sul-versao-final-2.pdf>. [accessed 13 March 2025].
 - [5] L. Tenório Filho, F. R. L. Mota, and O. M. D. B. Santana. Rede nacional de dados e os desafios da interoperabilidade entre os sistemas de informação na rede de atenção do sistema único de saúde no Brasil. *Revista Fontes Documentais*, 6(ed. especial):27–28, 2024. URL <https://periodicos.ufba.br/index.php/RFD/article/view/59913>. [accessed 7 January 2026].
 - [6] M. V. da S. Costa et al. Avanços e desafios da interoperabilidade no Sistema Único de Saúde. *Journal of Health Informatics*, 17(1):1112, 2025. doi: 10.59681/2175-4411.v17.2025.1112.
 - [7] Brasil. Ministério da Saúde. RNDS – Rede Nacional de Dados em Saúde, 2026. URL <https://rn.ds.saude.gov.br>. [accessed 7 January 2026].
 - [8] T. M. Braga, L. M. Soares, A. C. G. Ferreira, A. C. F. L. de Siqueira, J. M. V. de M. W. Held, M. E. C. N. da Gama, A. B. Frizon, R. B. Ribeiro, G. L. P. de Medeiros, and L. M. Silvestri. O uso de inteligência artificial na interpretação de exames médicos. *Brazilian Journal of Health Review*, 7(3):e70932, 2024. doi: 10.34119/bjhrv7n3-507.
 - [9] K. R. V. Leite, M. A. B. Pasquali, T. A. S. N. da Silva, L. V.-R. V. Rodrigues, G. D. X. Melo, A. P. Q. D. F. Pacheco, M. C. Lourenço Filho, J. R. R. Lopes, S. de S. Bezerra, M. E. F. Scaravonatti, M. C. D. de Sousa, F. B. de O. Torres, P. A. M. Guimarães, and J. B. C. P. dos Santos. Transformação da saúde: o impacto da inteligência artificial na medicina moderna. *Contribuciones a las Ciencias Sociales*, 17(5):e6726, 2024. doi: 10.55905/revconv.17n.5-086.
 - [10] M. A. Rahman, E. Victoros, J. Ernest, R. Davis, Y. Shanjana, and M. R. Islam. Impact of artificial intelligence (AI) technology in healthcare sector. *SAGE Open Medicine*, 12, 2024. doi: 10.1177/2632010X241226887.
 - [11] M. C. Oliveira et al. O uso da inteligência artificial na detecção precoce do câncer de mama: uma revisão de literatura. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, 10(10):3932–3940, October 2024. doi: 10.51891/rease.v10i10.16199.
 - [12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.
 - [13] Zafer Akcali et al. Automated extraction of key entities from non-English mammography reports using named entity recognition with prompt engineering. *Bioengineering*, 12(2):168, 2025. doi: 10.3390/bioengineering12020168.

- [14] Eduardo Godoy et al. Automatic detection of contextual laterality in mammography reports using large language models. In *Proceedings of the 14th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–6. IEEE, 2024. doi: 10.1109/ICPRS62101.2024.10677842.
- [15] Yongli Mou et al. Improving the quality of unstructured cancer data using large language models: A german oncological case study. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, volume 316 of *Studies in Health Technology and Informatics*, pages 685–689. IOS Press, 2024. doi: 10.3233/SHTI240507.
- [16] Yuxin Liu et al. Bootstrapping BI-RADS classification using large language models and transformers in breast magnetic resonance imaging reports. *Visual Computing for Industry, Biomedicine, and Art*, 8(1):1–16, 2025. doi: 10.1186/s42492-025-00189-8.
- [17] Dehua Chen et al. Extracting structure information from narrative medical reports based on LLMs. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5616–5623. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10822688.
- [18] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004. doi: 10.2307/25148625.
- [19] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007. doi: 10.2753/MIS0742-1222240302.
- [20] Vijay K. Vaishnavi and William Kuechler. *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. CRC Press, Boca Raton, 3 edition, 2021.
- [21] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.
- [22] LMArena. LMArena leaderboard, 2025. URL <https://lmarena.ai/leaderboard>. [accessed April 2025].
- [23] Yongjian Tang, Doruk Tuncel, Christian Körner, and Thomas A. Runkler. The few-shot dilemma: Over-prompting large language models. *arXiv preprint arXiv:2509.13196*, 2025. URL <https://arxiv.org/abs/2509.13196>.
- [24] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023. doi: 10.1162/tacl_a_00638.