Journal of Health Informatics

J Health Inform. 2026;18:e1847
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1847

ARTIGO COMPLETO
CBIS 2026

LuminaRad Sentinela: avaliação de sistema híbrido para detecção de erros em laudos radiológicos

LuminaRad Sentinela: evaluation of a hybrid system for detecting errors in radiology reports

Fernando Schaffazick[1] ✉, Laura Righi Boemo[2], Marcos Cordeiro D'Ornellas[3]

ORCID: 0009-0008-6883-4307 · 0009-0002-4748-1553 · 0000-0003-1883-3046

[1] Programa de Mestrado Profissional em Ciências da Saúde, Universidade Federal de Santa Maria (UFSM), Santa Maria (RS), Brasil.
[2] Universidade Federal de Santa Maria (UFSM), Santa Maria (RS), Brasil.
[3] Departamento de Computação Aplicada (DCOM), Universidade Federal de Santa Maria (UFSM), Santa Maria (RS), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1847

DOI
10.59681/2175-4411.v18.iEspecial.2026.1847

RECEBIDO
17/05/2026

APROVADO
20/07/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Fernando Schaffazick
feschaffa@gmail.com

LICENÇA

Este é um artigo publicado em acesso aberto sob a licença Creative Commons CC BY 4.0, que permite uso, distribuição e reprodução em qualquer meio, desde que o trabalho original seja corretamente citado.

CC BY 4.0

RESUMO

Objetivo: Avaliar a arquitetura híbrida de IA do módulo Sentinela, integrado à plataforma LuminaRad, para detecção em tempo real de erros em laudos radiológicos. **Métodos:** Arquitetura de três camadas (regras determinísticas, PLN com SpaCy e LLMs), avaliada com 100 laudos anonimizados (50 com erros/50 controles), comparando 10 LLMs quanto a F1-Score, latência e custo, e piloto de usabilidade com 10 profissionais (5 radiologistas e 5 residentes em Radiologia) da UDI/HUSM (~440 laudos reais). **Resultados:** regras atingiram 100% de acurácia em incongruência de sexo; PLN clássico apresentou 36% de falsos positivos; Gemini 2.5 Flash obteve F1-Score 0,911, latência 3,8s e custo US$0,003/requisição. A Escala de Usabilidade do Sistema (SUS) revelou pontuação 82,3 (grau A) e alta confiança percebida (4,7/5,0), mas 9 dos 10 avaliadores relataram falsos positivos ocasionais ou frequentes. **Conclusão:** em avaliação interna e piloto de produção, o módulo Sentinela/LuminaRad mostrou-se promissor e clinicamente viável na detecção de erros, equilibrando segurança do paciente e viabilidade operacional em ambiente hospitalar real.

Descritores: Inteligência Artificial; Radiologia; Segurança do Paciente.

ABSTRACT

Objective: To evaluate the hybrid AI architecture of the Sentinela module, integrated into the LuminaRad platform, for real-time error detection in radiology reports. **Methods:** A three-tier architecture (deterministic rules, NLP with SpaCy, and LLMs), evaluated with 100 anonymized reports (50 with errors/50 controls), comparing 10 LLMs on F1-Score, latency, and cost, plus a usability pilot with 10 professionals (5 radiologists and 5 radiology residents) at UDI/HUSM (~440 real reports). **Results:** rules achieved 100% accuracy for sex-incongruence errors; classical NLP showed 36% false positives; Gemini 2.5 Flash achieved F1-Score 0.911, 3.8s latency, and US$0.003/request. The System Usability Scale scored 82.3 (grade A) with high perceived confidence (4.7/5.0), but 9 of 10 evaluators reported occasional or frequent false positives. **Conclusion:** in this internal evaluation and production pilot, the Sentinela/LuminaRad module showed promise as a clinically viable tool for detecting errors, balancing patient safety and operational feasibility in a real hospital environment.

Keywords: Artificial Intelligence; Radiology; Patient Safety.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Artigo apresentado no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026), realizado de 23 a 25 de setembro de 2026, em Brasília (DF), e publicado na seção CBIS 2026 do Journal of Health Informatics.

COMO CITAR ESTE ARTIGO: Schaffazick F, Boemo LR, D'Ornellas MC. LuminaRad Sentinela: avaliação de sistema híbrido para detecção de erros em laudos radiológicos. J Health Inform 2026;18:e1847. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1847

1 INTRODUÇÃO

O laudo radiológico constitui o documento central de comunicação entre o médico radiologista e a equipe assistente, representando o registro oficial da interpretação de imagens médicas e a base sobre a qual decisões terapêuticas são tomadas [1,2]. A qualidade desse documento impacta diretamente a segurança do paciente: erros substanciais são estimados em até 4% na prática diária, podendo gerar diagnósticos tardios, tratamentos inadequados e implicações médico-legais graves [1,3]. Fatores como fadiga, alto volume de trabalho e o uso disseminado de softwares de reconhecimento de voz (SRS) agravam o problema, introduzindo novas categorias de erros (omissões, intrusões lexicais e falhas ortográficas), que frequentemente escapam à autoedição [4].

As abordagens tecnológicas disponíveis para detecção automática de erros são, em sua maioria, restritas a regras de correspondência de palavras-chave para problemas específicos, como lateralidade e gênero, carecendo de capacidade contextual [2,5]. A evolução do Processamento de Linguagem Natural (PLN) e dos Modelos de Linguagem Grandes (LLMs) representa uma mudança de paradigma: modelos como GPT-4 demonstraram capacidade comparável à de radiologistas na identificação de inconsistências, com vantagens de velocidade e custo [6,7]. Sistemas como AIspect e Nuance PowerScribe 360 avançaram nessa direção no contexto internacional, mas permanecem atrelados a ecossistemas específicos de sistemas de arquivamento e comunicação de imagens (PACS) e têm baixa penetração no mercado brasileiro [8].

No Brasil, a plataforma LuminaRad (www.luminarad.com.br) é uma iniciativa e desenvolvimento do primeiro autor, membro do Grupo CASA/UFSM, que atua como um "cinto de utilidades" para o radiologista, reunindo em uma única solução ferramentas de reconhecimento de voz avançado, auxílio por IA na formulação de laudos, *templates* estruturados e extensões para integração com sistemas PACS e RIS. A plataforma, incubada pela Pulsar/UFSM e já em operação comercial no modelo SaaS, integra o módulo Sentinela: um mecanismo automatizado de detecção de inconsistências em laudos radiológicos em tempo real. O presente trabalho, desenvolvido no âmbito do Programa de Mestrado Profissional em Ciências da Saúde da UFSM, descreve a avaliação científica da arquitetura híbrida de IA desse módulo; aspectos de UX/IHC da extensão Web não são objeto deste artigo.

Nenhuma solução atual combina simultaneamente: (1) análise híbrida com velocidade de regras determinísticas e profundidade semântica de LLMs; (2) integração agnóstica ao sistema PACS via extensão de navegador; e (3) feedback em tempo real durante a redação. Esta lacuna motivou o desenvolvimento e a avaliação do módulo Sentinela, cujo objetivo é avaliar a eficácia de uma arquitetura híbrida de IA para detecção de erros em laudos radiológicos em ambiente de produção real, contribuindo para a segurança do paciente e a eficiência do fluxo de trabalho do radiologista. O escopo avaliado limita-se a três categorias de erro: lateralidade e contradição/incongruência lógica (via LLM) e incongruência de sexo (via regras determinísticas); categorias interpretativas mais amplas, como omissão de achados, não são cobertas e permanecem como trabalho futuro.

Esta investigação apresenta três contribuições originais ao campo: (1) o primeiro *benchmarking* comparativo de grande escala de LLMs modernos (incluindo modelos das famílias GPT-5, Gemini 2.5 e Claude) na detecção de erros em laudos radiológicos em língua portuguesa; (2) a avaliação da arquitetura híbrida em ambiente de produção clínica real, com laudos genuínos de tomografia computadorizada (TC), ressonância magnética (RM) e radiografias no contexto hospitalar universitário; e (3) a avaliação integrada de viabilidade clínica, operacional e econômica de um módulo de detecção de erros integrado a uma plataforma comercial em operação.

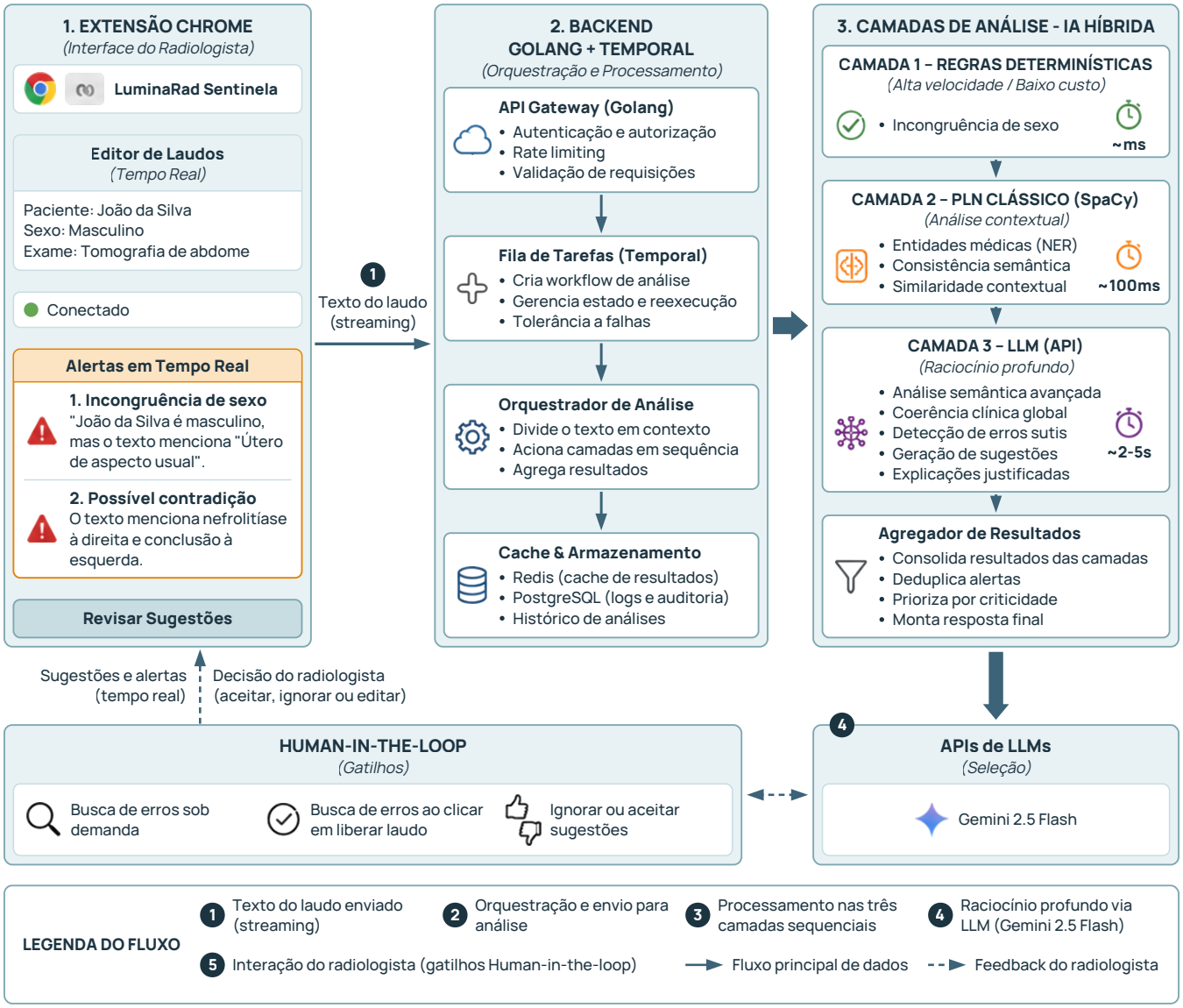
2 MÉTODOS

Este estudo descreve o desenvolvimento e a avaliação de um módulo de Suporte à Decisão Clínica (SDC) integrado à plataforma LuminaRad para detecção de erros em laudos radiológicos. O projeto foi submetido à avaliação da Comissão Científica da Gerência de Ensino e Pesquisa do Hospital Universitário de Santa Maria (HUSM) e, após, ao Comitê de Ética em Pesquisa do Hospital Universitário de Santa Maria (CEP-HUSM), reconhecido pela Comissão Nacional de Ética em Pesquisa (CONEP), sob o número CAAE 98831926.3.0000.0446, aprovado pelo parecer CEP nº 8.530.447. O projeto foi desenvolvido em conformidade com a Resolução do Conselho Nacional de Saúde (CNS) nº 466/2012. Por tratar-se de estudo observacional, retrospectivo, com dados obtidos a partir de laudos previamente adquiridos e devidamente anonimizados, foi concedida dispensa do Termo de Consentimento Livre e Esclarecido (TCLE) para a análise dos laudos. Os radiologistas participantes do estudo de usabilidade assinaram TCLE específico.

2.1 Contexto: Plataforma LuminaRad e Módulo Sentinela

O LuminaRad, plataforma SaaS descrita na Introdução, está disponível como aplicativo web (Vue.js), extensão para navegadores Chrome e Firefox e aplicativo móvel (Flutter). O módulo Sentinela, objeto do presente estudo, corresponde ao componente de detecção de erros da plataforma, operando de forma não intrusiva durante a redação do laudo.

Figura 1. Arquitetura de três camadas do módulo Sentinela integrado à plataforma LuminaRad.



Fonte: Elaborado pelos autores (2026).

2.2 Arquitetura do Sistema

O módulo Sentinela foi implementado como uma arquitetura cliente-servidor com três camadas analíticas complementares. A **Camada 1 (Regras Determinísticas)** realiza verificação instantânea de inconsistências estruturais, incluindo erros de lateralidade (discordâncias entre achados e conclusão) e incoerências sexo-específicas (menção de estruturas anatômicas incompatíveis com o sexo do paciente, inferido dos metadados do PACS). A **Camada 2 (PLN Clássico)** aplica análise sintática via SpaCy para identificação de relações complexas entre termos. A **Camada 3 (LLMs)** realiza análise semântica profunda do laudo completo via API, utilizando *prompt* padronizado com retorno estruturado em JSON. A integração com LanguageTool permite verificação ortográfica e gramatical em tempo real, com suporte a dicionário pessoal configurável e dicionário radiológico comunitário.

O *prompt* padronizado incluiu instrução de sistema definindo o papel do modelo como especialista em radiologia clínica, o texto integral do laudo e solicitação de retorno em JSON com um *array* de objetos, cada objeto representando um erro detectado, com dois campos: categoria (lateralidade ou contradição lógica) e um *array* de frases contraditórias. O mesmo *prompt* foi aplicado identicamente aos 10 modelos avaliados, garantindo comparabilidade; o texto completo está disponível mediante solicitação ao autor correspondente.

O *frontend* foi desenvolvido como extensão para o Google Chrome (JavaScript, Vue.js/Quasar, Manifest V3), publicada na Chrome Web Store. O *backend* foi implementado em Golang, explorando o modelo de concorrência de Goroutines para múltiplas requisições simultâneas com baixa latência. O Temporal atua como orquestrador de filas duráveis, garantindo retentativas automáticas e persistência de estado em caso de falhas da API. Toda a comunicação entre cliente e servidor

utiliza HTTPS, sem transmissão de dados identificadores do paciente, em alinhamento com os princípios de minimização de dados da LGPD [9]; a adequação contratual plena (retenção, uso para treinamento e transferência internacional pelos provedores de LLM) depende de avaliação institucional, discutida na Seção 4.

Durante o piloto (out. a nov. 2025), medidas adicionais de privacidade incluíram: processamento das regras determinísticas e do PLN inteiramente no *backend* local, sem transmissão externa; remoção automatizada de metadados identificadores (nome, data de nascimento, prontuário) antes do envio às APIs de LLM via HTTPS; e ausência de retenção persistente de laudos ou *logs* de conteúdo clínico além do processamento síncrono da requisição.

2.3 Fluxo "Human-in-the-Loop"

Para equilibrar custo operacional e segurança clínica, a verificação por LLM é acionada por dois gatilhos: o botão "Verificar inconsistências", disponível durante a redação, e o disparo automático ao "Liberar laudo". A verificação por regras ocorre de forma síncrona em tempo real, com pequeno atraso após o usuário parar de digitar. Quando inconsistências são detectadas na liberação, um modal não obstrutivo apresenta um resumo dos erros identificados, preservando a autonomia do médico para revisar ou prosseguir. Essa abordagem *Human-in-the-loop* foi adotada para mitigar a "fadiga de alertas", documentada como principal causa de desobediência a sistemas de apoio à decisão clínica [10,11].

2.4 Avaliação das Metodologias de Detecção

Para a comparação dos LLMs, utilizou-se um conjunto de avaliação de 100 laudos anonimizados da Unidade de Diagnóstico por Imagem do Hospital Universitário de Santa Maria (UDI/HUSM): 50 com erros propositalmente inseridos ou reais previamente identificados e 50 laudos corretos (controle). A distribuição 50/50 foi adotada intencionalmente para maximizar a capacidade discriminativa entre modelos, e não para estimar a prevalência real de erros no serviço (estimada em ~4% [1]). A referência interna foi estabelecida pelo primeiro autor, médico residente do terceiro ano de Radiologia e Diagnóstico por Imagem; a ausência de um segundo avaliador independente com cálculo de Kappa constitui limitação reconhecida do estudo, discutida na Seção 4. Os erros avaliados foram: (a) erros de lateralidade: discordâncias entre achados e conclusão; e (b) erros de congruência/contradição: afirmações logicamente incompatíveis (ex.: menção de colecistectomia prévia com vesícula descrita como presente). Dez modelos foram comparados quanto a F1-Score, Precisão, Recall, latência e custo por requisição.

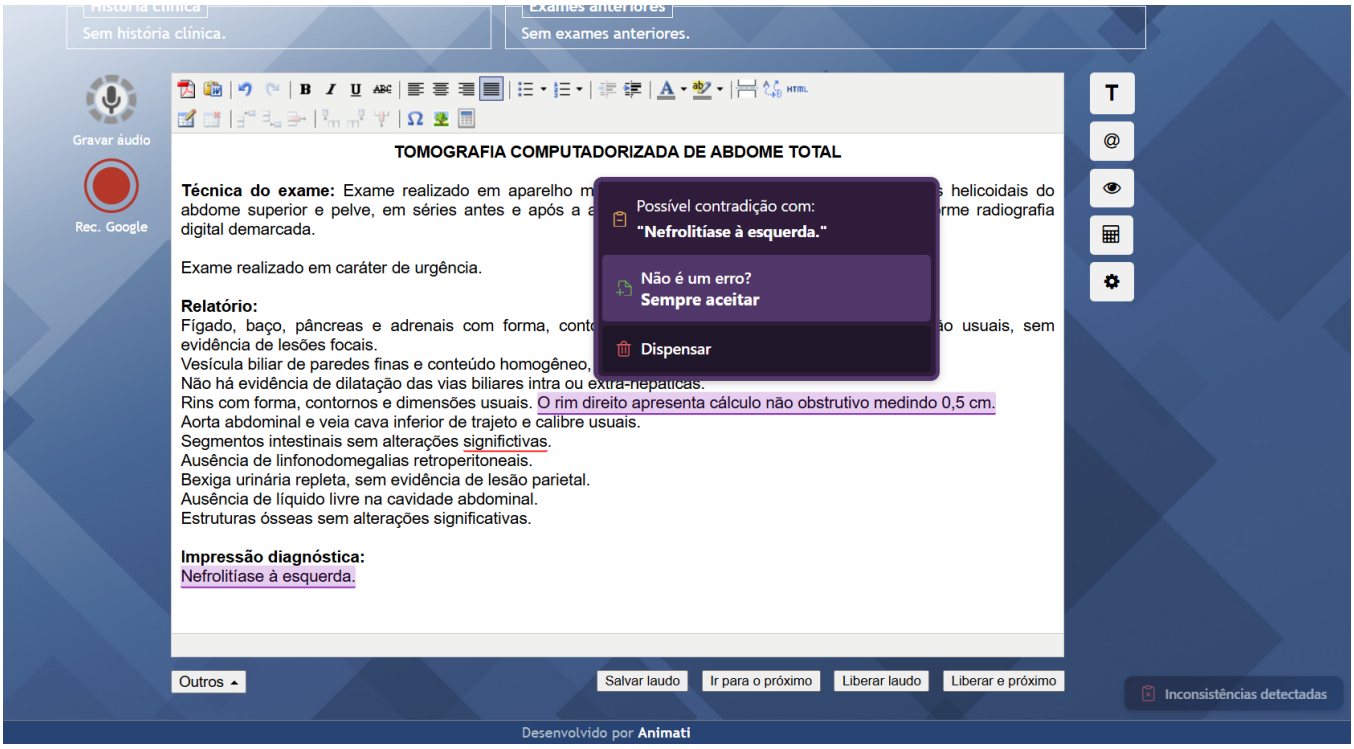
Os dez modelos foram selecionados por representatividade e disponibilidade comercial: à época (out. 2025), eram os modelos *flagship* ou de maior adoção de cada família: Google (Gemini 2.5 Pro, Flash, Flash Lite), OpenAI (GPT-5, GPT-5 Mini, GPT-5 Nano), Anthropic (Claude Sonnet 4.5, Claude 3.5 Haiku) e Mistral (Medium 3, Small 3.1). As variantes de cada família permitiram comparar o *trade-off* desempenho-custo-latência em diferentes faixas de capacidade, informando a seleção do modelo mais adequado à produção clínica.

O conjunto de 50 laudos com erros foi estruturado para incluir proporções representativas de ambas as categorias avaliadas: erros de lateralidade (discordâncias entre achados e conclusão do laudo) e erros de congruência/contradição lógica (afirmações clinicamente incompatíveis dentro do mesmo documento). Para os 50 laudos controle, verificou-se previamente a ausência de erros nas categorias avaliadas. A referência interna foi documentada em planilha estruturada com identificação do tipo de erro, trecho afetado e fundamentação clínica para cada caso, garantindo rastreabilidade das decisões do avaliador.

2.5 Avaliação de Usabilidade

A usabilidade foi avaliada durante implementação piloto na UDI/HUSM entre outubro e novembro de 2025, com análise de aproximadamente 440 laudos de TC, RM, radiografias convencionais (RX) e ultrassonografias (US). Participaram 10 profissionais (5 radiologistas e 5 residentes em Radiologia), selecionados por conveniência. O instrumento aplicado foi um questionário estruturado composto por: perfil do avaliador; a Escala de Usabilidade do Sistema (*System Usability Scale*, SUS), instrumento validado de 10 itens em escala de 1 a 5 [12]; questões estruturadas sobre relevância clínica e precisão percebida dos alertas, incluindo a percepção de falsos positivos; questões sobre impacto no fluxo de trabalho; e um campo de comentários livres. As respostas qualitativas abertas, obtidas de apenas 2 dos 10 participantes, foram tratadas de forma ilustrativa dada a baixa taxa de resposta, e não como resultado de uma análise temática indutiva formal.

Figura 2. Interface do módulo Sentinela/LuminaRad integrado ao PACS do HUSM (desenvolvido por Animati) durante uso em produção.



Fonte: Elaborado pelos autores (2026).

3 RESULTADOS

3.1 Implementação e Integração

O módulo Sentinela foi implementado com sucesso e integrado à plataforma LuminaRad. A extensão *standalone* está disponível na Chrome Web Store (URL: https://chromewebstore.google.com/detail/sentinela-assistente-de-l/eannapaahjikhgpldakmmmhdhpjholmg), bem como a extensão do LuminaRad (URL: https://chromewebstore.google.com/detail/luminarad-extens%C3%A3o/aipfdemgojknoegnnghobeabgjjkdnnm). O *backend* em Golang demonstrou capacidade de processar múltiplas requisições simultâneas com latência mínima nas camadas determinísticas, essencial para a percepção de feedback em "tempo real". A Figura 2 apresenta a interface do módulo integrado ao ambiente de laudamento do HUSM.

A captura de tela demonstra o destaque de erro de contradição entre informação contida no corpo do laudo e na conclusão. Ao manter o ponteiro do mouse sobre o erro, uma pequena janela é exibida explicando o porquê de aquela informação ter sido destacada. Além disso, é possível observar erro ortográfico marcado e pequeno indicador no canto inferior direito da tela mostrando que existem inconsistências no laudo.

3.2 Desempenho das Abordagens de Detecção

A abordagem baseada em regras determinísticas atingiu 100% de acurácia na detecção de erros de incoerência de sexo, com zero falsos positivos e latência submilissegundo, viabilizando feedback em tempo real. Isso se dá porque há um número limitado de palavras e estruturas pertencentes a um sexo ou outro (útero, próstata etc.). A aplicação de PLN clássico (SpaCy) para contradições semânticas resultou em taxa de falsos positivos de 36%, mesmo após diversas iterações de refinamento. Como taxas superiores a 15–20% comprometem a confiança do usuário pelo fenômeno de "fadiga de alertas" [10,11], optou-se por não utilizar essa abordagem no produto final.

A Tabela 1 apresenta os resultados da avaliação comparativa dos 10 LLMs. O Gemini 2.5 Flash destacou-se como a escolha ideal, combinando alto desempenho (F1-Score de 0,911; Precisão de 95,8%; Recall de 86,8%), latência operável de 3,8s e custo de US$0,0027/requisição, cerca de 3,5 vezes menor que o Gemini 2.5 Pro, que liderou em F1-Score (0,923), mas apresentou latência de ~10s, incompatível com uso interativo. Cabe ressalvar que, dado o tamanho amostral de 53 instâncias de erro, a diferença de F1-Score entre Flash e Pro situa-se dentro da margem de incerteza estatística (F1 do Gemini 2.5 Flash: IC95% 0,86–0,95), não sendo os dois modelos estatisticamente distinguíveis nesse critério; a escolha recaiu sobre o Flash pelo equilíbrio operacional entre latência e custo entre modelos de desempenho estatisticamente equivalente, e não por superioridade absoluta de F1-Score.

Tabela 1. Comparativo de desempenho, latência e custo entre LLMs na detecção de erros em laudos radiológicos. Avaliação realizada em outubro de 2025.

Modelo (API string / out.2025)	F1-Score	Precisão	Recall	TP	FP	FN	Latência (ms)	Custo (US$)
gemini/gemini-2.5-pro	0,923	94,1%	90,6%	48	3	5	9.812	$0,0096
gemini/gemini-2.5-flash	**0,911**	**95,8%**	**86,8%**	**46**	**2**	**7**	**3.805**	**$0,0027**
anthropic/claude-sonnet-4-5	0,907	100,0%	83,0%	44	0	9	12.885	$0,0164
mistral/mistral-medium-3	0,887	97,7%	81,1%	43	1	10	1.291	$0,0006
openai/gpt-5-mini	0,865	82,8%	90,6%	48	10	5	9.951	$0,0015
openai/gpt-5-nano	0,854	95,3%	77,4%	41	2	12	5.574	$0,0004
anthropic/claude-3-5-haiku-latest	0,822	100,0%	69,8%	37	0	16	3.405	$0,0017
openai/gpt-5	0,804	76,3%	84,9%	45	14	8	9.638	$0,0100
gemini/gemini-2.5-flash-lite	0,789	73,8%	84,9%	45	16	8	4.520	$0,0004
mistral/mistral-small-3	0,568	100,0%	39,6%	21	0	32	794	$0,0001

Legenda: TP = Verdadeiros Positivos; FP = Falsos Positivos; FN = Falsos Negativos. Linha em negrito = modelo selecionado. Custo em dólares americanos (US$).

Fonte: Elaborado pelos autores (2026).

Tabela 2. Dados agregados da implementação piloto do módulo Sentinela/LuminaRad na UDI/HUSM (out. a nov. 2025).

Modalidade	Laudos processados	Erros detectados (total)	Taxa de aceitação das sugestões*
TC (Tomografia)	200	18	~72%
RM (Ressonância)	50	4	~68%
RX (Radiografia)	80	5	~65%
US (Ultrassonografia)	110	12	~63%
Total aproximado	440	39	~67% (média)

Nota: valores aproximados extraídos dos registros de uso em produção. *Considerou-se "sugestão aceita" aquela em que o radiologista alterou o trecho sinalizado antes da liberação do laudo, e "não aceita" aquela em que o laudo foi liberado sem alteração do trecho apontado; a taxa de aceitação é a razão entre sugestões aceitas e o total de sugestões emitidas. Esse registro automático não permite distinguir, isoladamente, alertas não aceitos por serem falsos positivos daqueles descartados por discordância clínica legítima; a taxa de aceitação não deve, portanto, ser interpretada como equivalente a verdadeiro/falso positivo.

Fonte: Elaborado pelos autores (2026).

Modelos menores, Mistral Small (Recall=39,6%) e Gemini 2.5 Flash Lite (Recall=84,9%), falharam predominantemente em erros de congruência/contradição, que exigem raciocínio clínico mais profundo; erros de lateralidade, mais estruturais, foram detectados com maior consistência mesmo por modelos compactos.

3.3 Verificação Ortográfica

A verificação ortográfica via LanguageTool (Seção 2.2), com dicionário pessoal e comunitário radiológico refinado continuamente pelos usuários, reduziu falsos positivos ligados à terminologia técnica, particularmente relevante dado o volume de termos específicos da prática radiológica brasileira.

3.4 Viabilidade Econômica

Projetando 10.000 laudos/mês, o custo operacional com o Gemini 2.5 Flash seria de ~US$27,20/mês (R$~136,00, cotação de out. 2025), fração ínfima do faturamento de um serviço de radiologia de médio porte. O custo por laudo é 3,5× menor que o do Pro e ~6× menor que o do Claude Sonnet 4.5, argumento concreto de viabilidade para gestores hospitalares.

3.5 Dados de Uso em Produção e Usabilidade

A Tabela 2 sintetiza os dados agregados do piloto em produção, por modalidade de exame.

A ferramenta obteve pontuação SUS média de 82,3 (DP≈9,1; mediana 83,8; mín. 60; máx. 90; n=10), conceito "A" e adjetivo "Excelente" na escala de Bangor, Kortum e Miller [13], acima da média de referência de 68 obtida por Sauro [14] em mais de 5.000 estudos (percentil ~87). O item de confiança ("Eu me senti confiante ao usar o sistema") obteve média de 4,7±1,0.

Quanto à relevância clínica, 8/10 avaliadores classificaram a detecção de erros de sexo como "Excelente" e 2 como "Bom"; para inconsistências semânticas, 6 avaliaram como "Muito útil" e 4 como "Útil". Sobre falsos positivos, porém, apenas 1 avaliador os percebeu raramente, 6 ocasionalmente e 3 frequentemente, indicando que a taxa de falsos positivos permanece ponto de atenção relevante (Seção 4). Quanto ao fluxo de trabalho, 7 avaliadores classificaram a latência como "lenta, mas aceitável" e 3 como "rápida"; a verificação automática na liberação foi vista como essencial (n=4), útil (n=4) ou indiferente (n=2); e 6 relataram que o tempo de laudo não se

alterou, contra 4 com leve aumento compensado pela segurança percebida; nenhum relatou redução.

Apenas 2 dos 10 participantes responderam aos campos qualitativos abertos: um residente destacou a detecção de lateralidade e contradições como clinicamente relevante, apontando os falsos positivos como principal dificuldade de uso; outro relatou como ponto negativo a sinalização de trechos que optou por não alterar antes da liberação, ilustrando a dificuldade prática de distinguir falsos positivos de discordância clínica legítima. Dada a baixa taxa de resposta (2 de 10), esses relatos têm caráter ilustrativo, não configurando uma análise temática indutiva formal. Nenhum participante relatou cenários em que optaria por não utilizar a ferramenta.

4 DISCUSSÃO

Os resultados corroboram a premissa central do estudo: arquiteturas determinísticas combinadas a LLMs superam as abordagens tradicionais. A acurácia de 100% das regras para erros de sexo confirma os achados de Lee, Yang e Suh [15] e Min *et al.* [2] sobre a viabilidade de regras por palavras-chave em casos estruturais; já os 36% de falsos positivos do PLN clássico confirmam empiricamente sua insuficiência para inconsistências lógicas complexas, como antecipado por Min *et al.* [2] e Rao *et al.* [5].

A integração de LLMs superou essa barreira: o desempenho do Gemini 2.5 Flash (F1=0,911) está alinhado a Gertz *et al.* [6], que relataram 82,7% de detecção para o GPT-4 (150 erros inseridos em 200 laudos), comparável aos radiologistas do estudo, embora o leitor mais experiente tenha atingido quase 95%, e a Wu *et al.* [7], cujo LLM multimodal ajustado superou o especialista de referência em 1,67 p.p. no MIMIC-CXR. Esses números situam, a partir da literatura, a ordem de grandeza do ganho esperado com LLMs, ainda que a comparação direta entre o Sentinela e radiologistas sem auxílio computacional (não realizada aqui) permaneça necessária para quantificar esse ganho, como discutido adiante. A seleção do modelo considerou não apenas o F1-Score isolado, cuja diferença entre Flash e Pro não é estatisticamente significativa dado o tamanho amostral (Seção 3.2), mas o *trade-off* entre desempenho (F1-Score de 0,911 vs. 0,923 do Pro), latência (3,8s vs. ~10s) e custo (US$0,0027 vs. US$0,0096). Modelos compactos falharam no Recall de erros complexos (Mistral Small: 39,6%), evidenciando que o domínio clínico exige capacidade mínima de raciocínio não atingível por modelos menores.

O campo de LLMs avança rapidamente: os modelos avaliados refletem o estado da arte disponível em out. 2025, e gerações subsequentes tendem a ampliar ainda mais o desempenho e reduzir latências. Graças à arquitetura *plug-and-play*, o Sentinela pode incorporar esses ganhos apenas atualizando a chamada de API, sem reengenharia, o que recomenda monitoramento e reavaliação periódica do modelo em produção.

A avaliação em ambiente de produção real (com laudos clínicos de rotina da UDI/HUSM, não apenas simulados) diferencia este trabalho dos estudos anteriores baseados exclusivamente em *corpora* artificiais [2,5,16]. A taxa média de aceitação das sugestões (~67%), observada transversalmente às modalidades TC, RM, RX e US, indica que o sistema gera alertas com relevância clínica suficiente para motivar intervenção do radiologista na maioria dos casos.

Cabe qualificar a robustez das métricas da Tabela 1. Como a referência interna foi definida por um único avaliador, que também refinou iterativamente o *prompt* sobre o mesmo conjunto de 100 laudos, sem separação treino/teste (*holdout*), os valores de F1-Score, Precisão e Recall devem ser lidos como desempenho de avaliação interna, potencialmente otimista, e não como estimativa definitiva de produção. A ausência de um segundo avaliador com cálculo de Kappa impede quantificar o quanto a subjetividade de um único julgador influenciou casos-limite, sobretudo em erros de congruência/contradição, que exigem mais interpretação clínica que os de lateralidade. A taxa de aceitação no piloto (~67%), obtida sobre laudos inéditos e independentes do ajuste do *prompt*, oferece uma estimativa externa que mitiga parcialmente essa limitação, ainda que, como definido na Seção 3.5, não distinga falsos positivos de discordância clínica legítima, o que impede sua leitura direta como medida de acurácia.

Relacionada a essa ressalva está a distribuição artificial 50/50 do conjunto de avaliação, que não deve ser confundida com a prevalência real de erros no serviço (~4%). Projetando o Gemini 2.5 Flash para essa prevalência, a partir da sensibilidade de 86,8% (por instância) e da especificidade de 96,0% (por laudo-controle), o valor preditivo positivo (VPP) estimado reduz-se para ~47% (IC95% 38–56%): a cada 1.000 laudos, esperar-se-iam ~35 alertas verdadeiros e 38 falsos positivos. Precisão e Recall isolados, calculados sobre o conjunto balanceado, podem assim induzir leitura otimista do desempenho em produção; o VPP reduzido reforça a pertinência da verificação sob demanda e sob revisão humana obrigatória (Seção 2.3), que mitiga o risco de fadiga de alertas mesmo diante dessa carga esperada de falsos positivos.

A abordagem *Human-in-the-loop* mostrou-se assertiva para mitigar interrupções obstrutivas, refletindo-se na alta pontuação SUS (82,3; grau “A”) e na alta confiança relatada no item específico da escala (4,7/5,0). Ainda assim, a percepção de falsos positivos permaneceu elevada (apenas 1 dos 10 avaliadores relatou percebê-los raramente), consistente com a literatura sobre resistência a ferramentas com alta taxa de alarmes falsos [10,11] e reforçando a necessidade de revisão humana obrigatória mesmo diante de boa aceitação

geral. A implementação como extensão de navegador *plug-and-play* permite a transição da detecção de erros de auditoria retrospectiva para suporte interativo durante a redação, sem exigir adesão a ecossistemas específicos de PACS.

Uma limitação metodológica relevante é a ausência de comparação direta com a taxa de detecção de erros por radiologistas humanos sem auxílio computacional, o que teria permitido quantificar o ganho incremental de segurança oferecido pelo sistema. Tal comparação constitui objetivo prioritário de estudos futuros, nos quais o desempenho do conjunto radiologista + Sentinela será contrastado com o desempenho humano isolado, mensurando a redução de erros não identificados quando a ferramenta é utilizada como apoio à decisão clínica.

As principais limitações deste estudo incluem: (1) caráter unicêntrico, restrito à UDI/HUSM, limitando a generalização dos resultados; (2) referência interna estabelecida por um único avaliador (residente de 3º ano), sem cálculo de acordo interavaliadores (Kappa), o que reduz parcialmente a credibilidade formal do conjunto de avaliação; (3) distribuição artificial 50/50 no conjunto de avaliação, que pode superestimar os indicadores de desempenho em relação ao cenário clínico real (~4% de prevalência); (4) avaliação de usabilidade com amostra reduzida (n=10); (5) cobertura restrita a dois tipos de erro na detecção por LLM (lateralidade e contradição lógica); e (6) a verificação de incongruência de sexo, baseada em correspondência determinística com os metadados do PACS, não contempla cenários clínicos particulares, como pacientes pós-histerectomia ou pós-prostatectomia, transplantados, ou casos de incongruência entre sexo registrado e identidade de gênero. Estudos multicêntricos com amostras maiores, avaliadores independentes e espectro ampliado de categorias de erros são necessários para confirmar e estender os resultados obtidos.

Quanto à expansão para outros contextos do Sistema Único de Saúde, a arquitetura agnóstica ao PACS e de baixo custo computacional (Seções 2.2 e 3.4) favorece, em princípio, a portabilidade para hospitais de menor porte e regiões com infraestrutura limitada. A dependência de conectividade contínua e de APIs de LLM hospedadas no exterior, contudo, é uma barreira potencial em localidades com rede instável, o que sugere explorar futuramente modelos *open-source* auto-hospedados. Essa avaliação permanece especulativa, já que o estudo foi conduzido em um único centro de referência, não testado em unidades de menor porte.

5 CONCLUSÕES

O módulo Sentinela, integrado à plataforma LuminaRad, demonstrou que a arquitetura híbrida (regras determinísticas, PLN e LLMs) é promissora e clinicamente viável para detecção de erros em laudos radiológicos em tempo real. Principais achados: (1) regras com 100% de acurácia para erros de sexo; (2) PLN clássico insuficiente (36% de falsos positivos); (3) Gemini 2.5 Flash como escolha ótima no *trade-off* desempenho/latência/custo; e (4) taxa de aceitação de ~67% em produção real. A alta aceitação sugere potencial de integração ao fluxo clínico, com possível contribuição à segurança do paciente e à redução da carga cognitiva percebida, efeitos não mensurados como desfechos clínicos, o que caracteriza este estudo como avaliação interna e piloto de viabilidade, não validação clínica independente. Trabalhos futuros incluem estudos multicêntricos com validação prospectiva independente, segundo avaliador com Kappa, ampliação das categorias de erro, extensão a outros idiomas/especialidades e reavaliação contínua frente a novas gerações de LLMs.

Contribuição dos autores. Fernando Schaffazick: desenvolvimento do backend e do motor de detecção de erros do sistema LuminaRad Sentinela; coleta, análise e interpretação dos dados; redação do manuscrito. Laura Righi Boemo: desenvolvimento e aperfeiçoamento do frontend e da extensão para navegadores. Marcos Cordeiro D'Ornellas: orientação e supervisão científica da pesquisa; revisão crítica do manuscrito.

Agradecimentos. À Unidade de Diagnóstico por Imagem (UDI) do Hospital Universitário de Santa Maria (HUSM/UFSM) e a todos os radiologistas e residentes que participaram voluntariamente da avaliação piloto do sistema. À Pulsar Incubadora (UFSM) pelo suporte ao desenvolvimento da plataforma LuminaRad.

Financiamento. Esta pesquisa não recebeu financiamento específico de agências públicas, comerciais ou de organizações sem fins lucrativos.

Disponibilidade de dados e materiais. Os dados clínicos utilizados neste estudo não estão disponíveis publicamente em razão de restrições éticas e de confidencialidade. Informações agregadas e materiais metodológicos poderão ser disponibilizados pelo autor correspondente mediante solicitação fundamentada, observadas as autorizações institucionais aplicáveis.

Conflitos de interesse. O primeiro autor, Fernando Schaffazick, é idealizador, desenvolvedor e responsável técnico da plataforma comercial LuminaRad, à qual pertence o módulo Sentinela avaliado neste estudo, caracterizando potencial conflito de interesses de natureza comercial. Os demais autores não possuem vínculo comercial com a plataforma, conforme declarado no manuscrito.

Declaração de uso de IA. Este manuscrito foi redigido com auxílio de ferramenta de Inteligência Artificial Generativa (Claude, Anthropic) para apoio à estruturação do texto, revisão de linguagem e organização do conteúdo. Os dados, análises, resultados, conclusões e avaliações científicas são de inteira responsabilidade dos autores, que revisaram e aprovaram o conteúdo final em sua totalidade.

REFERÊNCIAS

1. Srinivasa Babu A, Brooks ML. The malpractice liability of radiology reports: minimizing the risk. Radiographics. 2015;35(2):547-54. doi: https://doi.org/10.1148/rg.352140046
2. Min D, Kim K, Lee JH, Kim Y, Park CM. RRED : A Radiology Report Error Detector based on Deep Learning Framework. In: Proceedings of the 4th Clinical Natural Language Processing Workshop. Stroudsburg, PA, USA: Association for Computational Linguistics; 2022. p. 41-52. doi: https://doi.org/10.18653/v1/2022.clinicalnlp-1.5
3. Ringler MD, Goss BC, Bartholmai BJ. Syntactic and semantic errors in radiology reports associated with speech recognition software. Health Informatics J. 2017;23(1):3-13. doi: https://doi.org/10.1177/1460458215613614
4. Minn MJ, Zandieh AR, Filice RW. Improving Radiology Report Quality by Rapidly Notifying Radiologist of Report Errors. J Digit Imaging. 2015;28(4):492-8. doi: https://doi.org/10.1007/s10278-015-9781-9
5. Rao VM, et al. ReXErr: Synthesizing Clinically Meaningful Errors in Diagnostic Radiology Reports. arXiv. 2024 set. Disponível em: http://arxiv.org/abs/2409.10829
6. Gertz RJ, Dratsch T, Bunck AC, Lennartz S, Iuga AI, Hellmich MG, et al. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. Radiology. 2024;311(1):e232714. doi: https://doi.org/10.1148/radiol.232714
7. Wu J, et al. Exploring Multimodal Large Language Models for Radiology Report Error-checking. arXiv. 2024a. Disponível em: http://arxiv.org/abs/2312.13103
8. AISPect. AISPect: ambient AI reporting for radiologists [Internet]. [citado 2026 nov 4]. Disponível em: https://aispect.ai/
9. Brasil. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD) [Internet]. Diário Oficial da União. 2018 ago 15 [citado 2026 out 18]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm
10. Parasuraman R, Riley V. Humans and Automation: Use, Misuse, Disuse, Abuse. Hum Factors. 1997;39(2):230-53. doi: https://doi.org/10.1518/001872097778543886
11. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc. 2006;13(2):138-47. doi: https://doi.org/10.1197/jamia.M1809
12. Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. International Journal of Human–Computer Interaction. 2008;24(6):574-94. doi: https://doi.org/10.1080/10447310802205776
13. Bangor A, Kortum P, Miller J. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. J Usability Stud. 2009;4(3):114–23.
14. Sauro J. A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. Denver, CO: Measuring Usability LLC; 2011.
15. Lee YH, Yang J, Suh JS. Detection and Correction of Laterality Errors in Radiology Reports. J Digit Imaging. 2015;28(4):412-6. doi: https://doi.org/10.1007/s10278-015-9772-x
16. Ostmeier S, Xu J, Chen Z, Varma M, Blankemeier L, Bluethgen C, et al. GREEN: Generative Radiology Report Evaluation and Error Notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg, PA, USA: Association for Computational Linguistics; 2024. p. 374-90. doi: https://doi.org/10.18653/v1/2024.findings-emnlp.21

NOTAS EDITORIAIS

O Journal of Health Informatics é uma publicação científica mantida pela Sociedade Brasileira de Informática em Saúde, associação científica sem fins lucrativos dedicada ao desenvolvimento da Informática em Saúde e da Saúde Digital no Brasil. O processo editorial é conduzido com independência científica, sem interferência institucional, comercial ou de patrocinadores, e segue as boas práticas internacionais de ética em publicação científica.

Este artigo foi avaliado por pares, em sistema duplo-cego, por especialistas da área. A equipe editorial usou ferramentas de inteligência artificial, com supervisão humana, na triagem inicial, na conferência do texto, das referências e dos números e, quando necessário, no redesenho de figuras, quadros e tabelas, aprovado pelos autores na prova. As ferramentas e as salvaguardas estão na Política de uso de inteligência artificial da revista. As decisões editoriais foram tomadas por pessoas.

SBIS Sociedade Brasileira de Informática em Saúde