J. Health Inform. 2025 Janeiro-Dezembro; 17

Artigo Original

DOI: https://doi.org/10.59681/2175-4411.v17.2025.1478

JHI JOURNAL OF HEALTH INFORMATICS

Otimização da auditoria interna de prontuários cirúrgicos: uma abordagem com IA

Optimization of internal auditing of surgical records: an AI-based approach

Optimización de la auditoría interna de historias clínicas quirúrgicas: un enfoque con IA

Rita de Cássia Almeida Sales[1], Isaura Romero Peixoto[2], Shirley da Silva Jacinto de Oliveira Cruz[3], Wellington Pinheiro dos Santos[4]

RESUMO

Objetivo: O objetivo foi desenvolver um assistente de auditoria hospitalar baseado em Inteligência Artificial Generativa, capaz de analisar e classificar informações dos prontuários cirúrgicos, para apoiar os procedimentos de auditoria interna no Hospital das Clínicas da Universidade Federal de Pernambuco (HC-UFPE). Métodos: A metodologia envolveu o treinamento de um modelo amplo de linguagem (LLM) com dados validados por especialistas, seguido da avaliação do seu desempenho por meio de sensibilidade, especificidade, índice Kappa e Área sob a curva ROC. A base de dados utilizada foi o Aplicativo de Gestão para Hospitais Universitários (AGHU), de onde os dados foram extraídos em formato json, após pré-processamento e tratamento no software Metabase. Foi adotado o ecossistema da OpenAI através da personalização do ChatGPT, modelo GPT-4º. Foi utilizada a técnica de prompt engineering, seguido pela técnica de fine-tuning no processo de treinamento. Resultados: Os resultados obtidos no desenvolvimento do assistente virtual para auditoria hospitalar demonstraram sensibilidade de 96,8%, especificidade de ≈ 77, índice Kappa de 0,78 e AUC ROC de 86,95%, indicando que o modelo é capaz de analisar prontuários com alta precisão. Os resultados demonstraram que o assistente consegue replicar eficazmente avaliações humanas, evidenciando potencial significativo para transformar a prática de auditoria hospitalar. Conclusão: conclui-se que o assistente de auditoria hospitalar pode otimizar processos do setor proposto, promovendo apoio, segurança, precisão e abrangência na auditoria de prontuários.

Descritores: Auditoria Hospitalar, Inteligência Artificial

ABSTRACT

Objective: The aim was to develop a Generative Artificial Intelligence–based hospital audit assistant capable of analyzing and classifying information from surgical medical records, in order to support internal auditing procedures at the Hospital das Clínicas of the Federal University of Pernambuco (HC-UFPE). Methods: The methodology involved training a large language model (LLM) with expert-validated data, followed by performance evaluation using sensitivity, specificity, Kappa index, and Area Under the ROC Curve (AUC-ROC). The database utilized was the Management Application for University Hospitals (AGHU), from which data was extracted in JSON format after pre-processing and treatment using Metabase software. The OpenAI ecosystem was adopted through the customization of ChatGPT, GPT-4º model. We applied prompt engineering followed by fine-tuning techniques during the training process. Results: The virtual assistant developed for hospital auditing demonstrated a sensitivity of 96.8%, specificity of approximately 77%, a Kappa index of 0.78, and AUC-ROC of 86.95%, indicating that the model effectively analyzes medical records with high precision. The results showed that the assistant could effectively replicate human evaluations, demonstrating significant potential for transforming hospital audit practices. Conclusion: It is concluded that the hospital audit assistant can optimize internal processes for the proposed sector, enhancing support, safety, accuracy, and comprehensiveness in medical record auditing.

Keywords: Hospital Audit, Artificial Intelligence

RESUMEN

Objective: The aim was to develop a Generative Artificial Intelligence–based hospital audit assistant capable of analyzing and classifying information from surgical medical records, in order to support internal auditing procedures at the Hospital das Clínicas of the Federal University of Pernambuco (HC-UFPE). Methods: The methodology involved training a large language model (LLM) with expert-validated data, followed by performance evaluation using sensitivity, specificity, Kappa index, and Area Under the ROC Curve (AUC-ROC). The database utilized was the Management Application for University Hospitals (AGHU), from which data was extracted in JSON format after pre-processing and treatment using Metabase software. The OpenAI ecosystem was adopted through the customization of ChatGPT, GPT-4º model. We applied prompt engineering followed by fine-tuning techniques during the training process. Results: The virtual assistant developed for hospital auditing demonstrated a sensitivity of 96.8%, specificity of approximately 77%, a Kappa index of 0.78, and AUC-ROC of 86.95%, indicating that the model effectively analyzes medical records with high precision. The results showed that the assistant could effectively replicate human evaluations, demonstrating significant potential for transforming hospital audit practices. Conclusion: It is concluded that the hospital audit assistant can optimize internal processes for the proposed sector, enhancing support, safety, accuracy, and comprehensiveness in medical record auditing.

Descriptores: Auditoría Hospitalaria, Inteligencia Artificial

[1] *Enfermeira, Mestranda em Engenharia Biomédica, Universidade Federal de Pernambuco, https://orcid.org/0000-0003-3133-8955*

[2] *Médica Geriatra do Hospital das Clínicas da Universidade Federal de Pernambuco, https://orcid.org/0000_0002_9256_5054*

[3] *Doutora em Ciência da Computação, Chefe do Setor de Tecnologia da Informação e Saúde Digitial do Hospital das Clínicas da UFPE, https://orcid.org/0000-0002-8289-5105*

[4] *Professor do Departamento de Engenharia Biomédica, Doutor em Engenharia Elétrica, Universidade Federal de Pernambuco, https://orcid.org/0000-0003-2558-6602*

Autor Correspondente: **Wellington Pinheiro dos Santos**
e-mail: **wellington.santos@ufpe.br**

Artigo recebido: 27/03/2025
Aprovado: 25/12/2025
https://jhi.sbis.org.br/

INTRODUÇÃO

A auditoria de prontuários hospitalares é uma atividade essencial para a manutenção da qualidade e conformidade dos registros assistenciais, sendo um pilar para a segurança do paciente e a eficiência dos serviços de saúde. Os registros assistenciais no prontuário do paciente representam a documentação detalhada dos diagnósticos, tratamentos, intervenções e a evolução clínica, permitindo que toda a equipe de saúde acompanhe o histórico do paciente e tome decisões mais justificadas. Além disso, esses registros promovem a comunicação entre profissionais, de forma a garantir a conformidade legal e ética dos serviços prestados, contribuindo para a melhoria contínua dos processos assistenciais, da auditoria e segurança do paciente[1-2].

A auditoria de prontuários hospitalares é um componente essencial da qualidade assistencial, pois assegura a integridade dos registros clínicos e o cumprimento das normas legais e éticas, além de subsidiar a gestão e a segurança do paciente. Os registros no prontuário do paciente constituem uma fonte primária de informação clínica, permitindo a rastreabilidade das condutas, a comunicação multiprofissional e a melhoria contínua dos processos de cuidado[1-2].

A auditoria hospitalar também exerce papel educativo e estratégico, ao identificar falhas e propor melhorias que fortalecem a eficiência institucional e a sustentabilidade financeira[3]. No entanto, a ausência de padronização dos registros e a predominância de processos manuais tornam a auditoria uma atividade morosa e suscetível a erros humanos[4-5].

A aplicação de Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) desponta como solução promissora para automatizar a análise de grandes volumes de informações textuais, possibilitando identificar padrões e inconformidades com rapidez e precisão. Em contextos clínicos, técnicas como machine learning, data mining e large language models (LLMs) têm sido exploradas para apoiar decisões assistenciais e administrativas[6].

Estudos internacionais demonstram avanços significativos: sistemas de NLP têm sido empregados para extração de achados clínicos, codificação automatizada e identificação de eventos adversos, como observado no Reino Unido e nos Estados Unidos[7-8-9]. Ainda assim, persiste uma lacuna no uso de modelos generativos de larga escala aplicados à auditoria hospitalar, especialmente em prontuários cirúrgicos e em língua portuguesa, dentro do contexto regulatório do Sistema Único de Saúde (SUS)[7-8].

Diante desse cenário, o presente estudo teve como objetivo desenvolver e validar um assistente de auditoria hospitalar baseado em Inteligência Artificial Generativa, capaz de analisar e classificar informações de prontuários cirúrgicos, apoiando os processos de auditoria interna do Hospital das Clínicas da Universidade Federal de Pernambuco (HC-UFPE). A proposta busca integrar técnicas de prompt engineering e fine-tuning em um modelo de linguagem de grande porte, explorando o potencial dessa tecnologia para aprimorar a eficiência, a precisão e a segurança das auditorias clínicas.

MÉTODOS

A presente pesquisa propôs desenvolver uma aplicação vertical utilizando Inteligência Artificial (IA), especificamente um Modelo Amplo de Linguagem (LLM), para apoiar processos de auditoria hospitalar interna. O propósito central foi automatizar e otimizar a análise de prontuários hospitalares, permitindo identificar conformidades e não conformidades com maior precisão e eficiência.

O modelo foi desenvolvido e estruturado conforme as seguintes fases: (1) Seleção de Dados, (2) Pré-processamento e preparação dos dados, (3) Processamento pelo LLM, (4) Validação da solução.

1. Seleção dos Dados

Na primeira fase, o objetivo foi identificar e selecionar os dados mais relevantes e disponíveis para auditoria hospitalar no formato digital. Utilizou-se como fonte principal o banco de dados do Aplicativo de Gestão para Hospitais Universitários (AGHU), sistema integrado e adotado por todos os 45 hospitais da Rede Ebserh, incluindo o Hospital das Clínicas da Universidade Federal de Pernambuco (HC-UFPE). Este sistema reúne informações clínicas e administrativas detalhadas sobre os pacientes, oferecendo dados em formatos estruturados (como tabelas e campos padronizados) e não estruturados (textos livres, evoluções clínicas).

Inicialmente, foi realizada uma definição do escopo informacional essencial à auditoria interna realizada pelos profissionais do Núcleo Interno de Auditoria em Saúde (NIAS). O foco do estudo foi em prontuários de pacientes submetidos a procedimentos cirúrgicos, pois estes concentram maior complexidade documental e regulatória para fins de auditoria.

Como critérios de Inclusão foram utilizados: prontuário do paciente no formato digital; prontuários de pacientes com registro de procedimentos cirúrgicos, ocorridos entre janeiro de 2023 e dezembro de 2024; e prontuários com volume de texto compatível com o limite de 8500 tokens.

Para garantir representatividade do subuniverso estudado, adotou-se uma amostra inicial de 30 prontuários para a validação do modelo após o treinamento por prompt engineering. Na etapa de fine-tuning, foi utilizada uma amostra adicional de 30 prontuários, dividida em 10 prontuários para treino e 20 para teste. Esta seleção foi realizada de forma aleatória dentre os prontuários que atenderam aos critérios de inclusão.

O NIAS realiza a auditoria de uma média de 500 prontuários mensalmente. A amostra final de prontuários representa aproximadamente 10% do total mensal auditado. É fundamental reconhecer que o tamanho amostral é limitado pela restrição técnica de processamento do modelo de LLM utilizado. Essa limitação impôs a inclusão apenas de prontuários de internações curtas, resultando na exclusão intencional de casos mais longos e complexos.

Embora esta restrição limite a generalização dos resultados para o universo completo de auditoria, ela garante que o processamento pelo LLM seja realizado com informações completas e não truncadas, sendo um passo metodológico essencial para validar a eficácia da abordagem em um ambiente controlado. O estudo, portanto, foca na viabilidade técnica da solução para um subconjunto específico de alta ocorrência (prontuários curtos), sendo as limitações decorrentes do tamanho e escopo da amostra discutidas em detalhes na seção de Discussão.

O presente estudo foi submetido e aprovado pelo Comitê de Ética em Pesquisa (CEP) do HC-UFPE, conforme o Parecer nº 7112744. Para conformidade com a Lei Geral de Proteção de Dados (LGPD) e as diretrizes éticas, empregou-se a técnica de supressão dos dados pessoais sensíveis, excluindo informações como nome completo, data de nascimento, filiação ou qualquer dado que pudesse identificar diretamente os pacientes.

2. Pré-processamento e Preparação dos Dados

Após seleção inicial, realizou-se o processo de extração e pré-processamento dos dados com a utilização do software Metabase, uma plataforma open-source de business intelligence conhecida pela facilidade em explorar, filtrar e visualizar dados de forma intuitiva.

Nesta fase, os dados foram submetidos a um processo de limpeza e normalização, incluindo remoção de informações redundantes ou inválidas, padronização terminológica e organização das informações conforme padrões pré-estabelecidos pela auditoria. Esse processo garantiu consistência, precisão e qualidade, características indispensáveis ao processamento eficiente e eficaz pelo modelo de inteligência artificial.

Para adequar-se ao limite técnico do modelo de IA disponível (8500 tokens por requisição), optou-se por priorizar registros cuja extensão de informações permitisse um processamento completo dos dados evitando o truncamento de informações críticas durante as análises automatizadas. Após esses procedimentos, os dados pré-processados foram convertidos e estruturados no formato JSON, apropriado para envio às interfaces (APIs) da OpenAI.

Considerando o contexto real da auditoria interna no HC-UFPE, a distribuição das classes de auditoria é inerentemente desbalanceada, com uma frequência de classes categorizadas como 'Conformes' significativamente maior do que aqueles classificados como 'Não Conformes'. Na amostra final de prontuários utilizada para fine-tuning e validação, a proporção observada foi de 75% Conformes para 25% Não Conformes. Este desbalanceamento não foi corrigido por técnicas como oversampling ou undersampling na fase de fine-tuning.

3. Processamento pelo Modelo de IA

Nesta etapa central do estudo, o processamento foi realizado utilizando o modelo GPT-4o, fornecido pela OpenAI, um sistema avançado baseado na arquitetura Transformer com capacidade de analisar textos complexos e captar contextos detalhados. Esse modelo possui um processamento estimado de 1,76 trilhão de parâmetros, conferindo-lhe excepcional desempenho em tarefas de compreensão textual e geração de respostas coerentes.

A interação com o modelo foi estabelecida via API da OpenAI, com dados estruturados em formato JSON. Foi adotada a técnica avançada conhecida como prompt engineering, um método que visa aprimorar significativamente a precisão, coerência e contexto das respostas geradas pelo modelo. Por meio de técnicas como meta prompting e chain of thought[6], buscou-se maximizar a capacidade interpretativa e analítica do modelo, induzindo respostas alinhadas às diretrizes regulatórias específicas da auditoria hospitalar.

Adicionalmente, para tornar o modelo ainda mais robusto e adequado ao domínio específico da saúde, foi criada e inserida uma base complementar de documentos normativos e regulatórios. Esta base incluiu resoluções dos Conselhos Federais de Medicina e Enfermagem, documentos oficiais sobre procedimentos clínicos e diretrizes detalhadas sobre auditoria hospitalar.

Em seguida, foi realizado um procedimento de treinamento supervisionado (Supervised Fine-Tuning) do modelo com o objetivo de refinar ainda mais a precisão das análises automatizadas. Para isso, foi utilizada uma amostra adicional de 30 prontuários, preparados com os mesmos critérios de inclusão. A divisão dos prontuários adotada para os treinamentos foi de 10 prontuários para treino e 20 para teste.

O modelo de auditoria foi treinado de forma supervisionada, utilizando o GPT-4o-2024-08-06 como base. O fine-tuning seguiu um protocolo reprodutível com seed 1118733444, batch size = 1, learning rate multiplier = 2.0 e 3 epochs, totalizando cerca de 721 mil tokens. O dataset continha interações reais entre auditores e o assistente de auditoria, formatadas em JSON. As configurações adotadas visaram maior precisão em um conjunto pequeno e textual, permitindo ajustes mais granulares (batch = 1) e convergência acelerada (LR × 2.0). O número limitado de epochs buscou equilibrar aprendizado e generalização.

Durante o treinamento, os logs registraram validação, checkpoints (~70 e 140 steps) e conclusão sem falhas. As curvas de perda e acurácia mostraram padrão serrilhado, porém estável, coerente com os hiperparâmetros. O modelo resultante foi identificado como ft:gpt-4o--2024-08-06:estrategia-educacional.

Como limitações, reconhece-se que o batch = 1 e a taxa de aprendizado elevada podem aumentar a variância, e 3 épochs trazem risco moderado de overfitting em bases pequenas. Recomenda-se a repetição com seeds distintas e análise de intervalos de confiança para comprovar robustez. Ainda assim, a estabilidade dos registros e a coerência das métricas atestam a consistência e reprodutibilidade do treinamento.

4. Validação da Solução

Na fase final, os resultados obtidos através das auditorias automatizadas realizadas pelo modelo foram validados mediante comparação direta com análises realizadas manualmente pelos auditores especialistas do Núcleo Interno de Auditoria em Saúde (NIAS) do HC-UFPE.

Primeiramente, testou-se o desempenho inicial do sistema utilizando os 30 prontuários selecionados. Após os ajustes do fine-tuning, realizou-se nova rodada de validação, utilizando um conjunto de 20 prontuários. Os resultados da auditoria automatizada foram comparados e quantificados frente às análises humanas para avaliar o grau de correspondência entre ambas.

As métricas estatísticas adotadas para validação do modelo foram sensibilidade, especificidade e área sob a curva ROC. Essas métricas são amplamente utilizadas em estudos que empregam modelos de aprendizado de máquina, permitindo uma avaliação precisa e criteriosa da capacidade do modelo em classificar corretamente registros conformes e não conformes.

Por fim, a saída final das auditorias realizadas pelo modelo foi padronizada em um relatório estruturado contendo as seguintes seções: Identificação do atendimento; Anamnese e evoluções médicas; Informações sobre cirurgias realizadas; Anamnese e evoluções de enfermagem; contagem da atuação da equipe multiprofissional; recomendações finais.

RESULTADOS E DISCUSSÃO

Na primeira etapa do estudo, que utilizou exclusivamente a técnica de prompt engineering, o modelo foi testado em 30 prontuários e comparado às análises dos auditores humanos do Núcleo Interno de Auditoria em Saúde (NIAS) do HC-UFPE. A matriz de confusão (Tabela 1) demonstra o desempenho inicial, com sensibilidade de 96,8%, especificidade de 57%, índice Kappa de 0,61 e AUC-ROC de 77%. Esses valores indicam que o modelo apresentou elevada capacidade de reconhecer registros conformes, mas dificuldade moderada na identificação dos não conformes, resultando em maior incidência de falsos positivos.

Tabela 1 - Matriz de confusão da etapa 1 do estudo

Especialista/ Assistente	Positivo (C)	Negativo (NC)
Positivo (C)	TP = 901	FN = 29
Negativo (NC)	FP = 129	TN = 171

Fonte: A autora (2025)

Note: TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative.

O resultado inicial confirmou a viabilidade técnica da solução, embora evidenciasse a necessidade de ajuste fino para reduzir a permissividade nas classificações. Na segunda fase, foi aplicado o fine-tuning supervisionado, utilizando 10 prontuários para treino e 20 para teste. A nova matriz de confusão (Tabela 2) revelou melhora expressiva da performance, com sensibilidade de 97,6%, especificidade de 76,3%, Kappa de 0,78 e AUC-ROC de 86,95%, demonstrando avanço substancial na precisão e na concordância com as análises humanas.

Os resultados obtidos no desenvolvimento do assistente virtual para auditoria hospitalar demonstraram um desempenho com uma sensibilidade de 96,8% indicando que o modelo é capaz de analisar prontuários com alta precisão.

Já a especificidade mostrou-se moderada, com índice de ≈ 57%, revelando que apesar de reconhecer bem os casos conformes, o modelo apresenta maior dificuldade em identificar corretamente os registros não conformes (NC).

O índice Kappa também foi razoável (~0,61). Ao considerar o acaso, o valor de Kappa indica um nível

de concordância moderado entre o GPT e os especialistas.

O valor da AUC ROC de ≈ 77% mostra que, de maneira global, o GPT possui boa capacidade para distinguir entre critérios conformes e não conformes, porém ainda não atinge um patamar de excelência. É um desempenho satisfatório, porém não ótimo, sugerindo que otimizações adicionais podem melhorar a diferenciação entre C (conforme) e NC (Não conforme).

No conjunto, esses resultados indicam que o modelo possui uma performance global boa, mas com potencial de aprimoramento no que diz respeito à identificação mais acurada das não conformidades.

Na segunda fase do estudo, foram utilizados 10 prontuários para realizar o fine-tuning do modelo, ajustando-o às particularidades do contexto hospitalar e às necessidades particulares da auditoria. A tabela 2 demonstra a matriz de confusão dessa nova etapa, ou seja, após aplicação do fine-tuning.

Tabela 2 – Matriz de confusão da segunda análise comparativa, ou seja, comparação entre o modelo após Fine-tuning e os especialistas

Especialista/ Assistente	Positivo (C)	Negativo (NC)
Positivo (C)	TP = 569	FN = 14
Negativo (NC)	FP = 56	TN = 181

Fonte: A autora (2025)

Note: TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative.

Após esse processo, o desempenho do assistente virtual foi testado nos 20 prontuários restantes da amostra, resultando em uma sensibilidade de 97,6%. Observa-se que não se tratou de um incremento robusto no desempenho. Todavia, a especificidade, apresentou melhora significativa, indicando que o modelo não apenas classifica corretamente as informações, mas também minimiza falsos positivos e falsos negativos.

A comparação entre as duas fases evidencia que, mesmo com uma amostra reduzida, o modelo refinado foi capaz de reproduzir de forma consistente o julgamento dos auditores, apresentando melhor equilíbrio entre sensibilidade e especificidade. Essa melhoria reflete o ganho semântico obtido com o fine-tuning, que permitiu ao modelo compreender melhor nuances contextuais de linguagem clínica — especialmente nas seções de anamnese e evolução, historicamente críticas para a auditoria hospitalar.

O aumento do coeficiente Kappa de 0,61 para 0,78 confirma maior concordância estatística entre o assistente e os especialistas, evidenciando um nível de desempenho comparável ao de métodos humanos de dupla checagem. O ganho de 10 pontos percentuais na AUC-ROC reforça o avanço do modelo na distinção entre registros conformes e não conformes, alinhando-se aos padrões observados em implementações internacionais de NLP clínico[7-8-11].

Apesar dos resultados promissores, foram identificados desafios típicos de modelos de linguagem de grande porte, como interpretação ambígua de contextos incompletos e ocorrência pontual de alucinações. Esses comportamentos, relatados em estudos recentes[12-13-14], refletem limitações na retenção de informações centrais em contextos extensos, fenômeno conhecido como lost in the middle[11-12]. No presente estudo, observou-se que o desempenho tendia a cair quando a informação-chave estava no meio do texto clínico, o que confirma o viés posicional descrito na literatura.

Para mitigar esses efeitos, recomenda-se o uso de estratégias complementares, como prompts com encadeamento lógico (chain-of-thought), feedback humano contínuo e validação cruzada por LLM secundário. Essas abordagens, já empregadas em sistemas de apoio à decisão clínica, reduzem a incidência de respostas incoerentes e aumentam a rastreabilidade das inferências geradas.

As principais limitações metodológicas concentram-se no tamanho e escopo da amostra, restrita por limites técnicos de processamento de 8.500 tokens por prontuário, o que levou à seleção de casos de curta permanência. Essa amostra representa cerca de 10% dos prontuários auditados mensalmente pelo HC-UFPE, assegurando representatividade inicial, mas limitando a generalização dos resultados.

O desbalanceamento natural entre classes conformes (75%) e não conformes (25%) também impôs restrições, embora o modelo tenha mostrado capacidade de identificar padrões minoritários com elevada sensibilidade.

Além disso, optou-se por não realizar comparação direta com baselines clássicos de aprendizado de máquina, como regressão logística ou random forest, uma vez que o foco deste trabalho foi avaliar a viabilidade do uso de LLMs generativos em auditorias de texto clínico não estruturado — cenário em que tais modelos apresentam vantagens inerentes de aprendizado contextual (few-shot learning). Futuros estudos deverão incorporar essa comparação para quantificar ganhos específicos de desempenho.

Essas observações estão em consonância com a literatura internacional, que destaca a necessidade de estudos de validação multicêntrica e análise de custo-efetividade para a adoção de NLP em auditoria clínica[15-16-17].

Em síntese, o assistente virtual desenvolvido apresentou robustez técnica e consistência estatística, mesmo em cenário controlado. As evidências indicam que a abordagem pode reduzir o tempo de auditoria, minimizar vieses humanos e fortalecer a rastreabilidade regulatória, mantendo a supervisão clínica como componente essencial. A consolidação dessa tecnologia requer novas fases de ampliação de dados e integração com o sistema AGHU, visando testar sua escalabilidade e impacto operacional na Rede Ebserh.

Implicações Futuras

Os resultados deste estudo indicam perspectivas concretas para a evolução da auditoria clínica assistida por IA. Pesquisas futuras devem explorar modelos de linguagem abertos com maior capacidade contextual, bem como técnicas de sumarização de prontuários que permitam abranger casos de longa permanência.

A realização de análises custo-benefício e o desenvolvimento de protocolos de integração com o sistema AGHU são etapas fundamentais para a escalabilidade da solução. Reforça-se, ainda, que o uso de estruturas híbridas de decisão — combinando automação e supervisão humana — constitui o caminho mais seguro e eficiente para a adoção ética e sustentável da IA em ambientes hospitalares.

CONCLUSÃO

O estudo demonstrou que, mesmo com uma amostra limitada, o modelo de auditoria baseado em Inteligência Artificial Generativa apresentou alto desempenho na classificação de prontuários cirúrgicos, com sensibilidade de 97,6%, especificidade de 76,3%, coeficiente Kappa de 0,78 e AUC-ROC de 86,95%. Esses resultados confirmam a capacidade do assistente virtual de reproduzir análises humanas com consistência estatística e precisão operacional.

As técnicas combinadas de prompt engineering e fine-tuning mostraram-se eficazes para adaptar o modelo de linguagem ao domínio clínico-regulatório, evidenciando o potencial da IA generativa como ferramenta de apoio às auditorias hospitalares. Apesar dos avanços, observou-se a necessidade de aprimorar a interpretação de contextos complexos e a mitigação de alucinações — aspectos que reforçam a importância da supervisão humana contínua no uso desses sistemas[17-18].

Em síntese, a abordagem proposta representa um avanço no uso de LLMs em auditoria de prontuários, contribuindo para maior eficiência, padronização e rastreabilidade dos processos. A integração futura com sistemas institucionais, como o AGHU, e a expansão para amostras multicêntricas permitirão avaliar a escalabilidade e o impacto real dessa tecnologia sobre a qualidade da informação e a segurança do paciente.

REFERÊNCIAS

1. Pinheiro MB, Campos RKGG, Maniva SJC de F, Rolim KMC, Bonfim IM. ecnologias disponíveis para o processo de auditoria interna em classificação de risco: revisão integrativa. Rev Bras Pesqui Em SaúdeBrazilian J Health Res. 2023;25(4):81–8.
2. Jacomini LDS, Mangiavacchi BM. AUDITORIA HOSPITALAR COMO INSTRUMENTO NA PREVENÇÃO DE ERROS MEDICAMENTOSOS. Múltiplos Acessos. 9 de junho de 2021;6(1):199–207.
3. Ceretta JC, Seibert RM, Callegaro ARC. Gestão hospitalar: a auditoria operacional como ferramenta estratégica para o controle de desperdícios. Rev Gest E Secr. 1o de março de 2023;14(3):2663–75.
4. Pimentel LCL, Virginio JPA, Albernaz CB, Souza DP de, Jesus CAC de, Paranaguá TT de B, et al. Avaliação da qualidade dos registros do processo de enfermagem por meio de auditoria retrospectiva. Rev Enferm UERJ. 12 de dezembro de 2023;31:e77316–e77316.
5. Caixeta FDC, Silva AA, Ferreira MV, Sales RDCA. AUDITORIA EM SAÚDE: NÃO CONFORMIDADES EM REGISTROS DE PRONTUÁRIOS CIRÚRGICOS. Int J Health Sci. 30 de outubro de 2024;3(2):181–2.
6. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT [Internet]. arXiv; 2023. [citado 25 de março de 2025] Disponível em: http://arxiv.org/abs/2302.11382
7. Wu H, Wang M, Wu J, Francis F, Chang YH, Shavick A, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. npj Digit Med. 2022;5:186. doi:10.1038/s41746-022-00730-6
8. Afshar M, Adelaine S, Resnik F, Mundt MP, Long J, Leaf M, et al. Deployment of real-time natural language processing and deep learning clinical decision support in the electronic health record: pipeline implementation for an opioid misuse screener in hospitalized adults. JMIR Med Inform. 2023;11:e44977. doi:10.2196/44977
9. Au Yeung J, Shek A, Searle T, Kraljevic Z, Dinu V, Ratas M, et al. Natural language processing data services for healthcare providers. BMC Med Inform Decis Mak. 2024;24:356. doi:10.1186/s12911-024-02713-x.
10. An S, Ma Z, Lin Z, Zheng N, Lou JG. Make Your LLM Fully Utilize the Context [Internet]. arXiv; 2024. [citado 25 de março de 2025] Disponível em: http://arxiv.org/abs/2404.16811
11. Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, et al. Lost in the Middle: How Language Models Use Long Contexts [Internet]. arXiv; 2023. [citado 25 de março de 2025] Disponível em: http://arxiv.org/abs/2307.03172
12. Baker GA, Raut A, Shaier S, Hunter LE, Wense K von der. Lost in the Middle, and In-Between: Enhancing Language Models' Ability to Reason Over Long Contexts in Multi-Hop QA [Internet]. arXiv; 2024. [citado 25 de março de 2025] Disponível em: http://arxiv.org/abs/2412.10079
13. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A Survey on Evaluation of Large Language Models [Internet]. arXiv; 2023. [citado 25 de março de 2025] Disponível em: http://arxiv.org/abs/2307.03109
14. Verspoor K. "Fighting fire with fire" - using LLMs to combat LLM hallucinations. Nature. junho de 2024;630(8017):569–70.
15. PERKINS, C. et al. The Economic and Clinical Burden of Documentation in Healthcare: A Call for AI-Powered Solutions. Journal of Clinical Documentation, v. 38, n. 2, p. 112-125, 2024.
16. EGUIA, Hans et al. Suporte à Decisão Clínica e Processamento de Linguagem Natural em Medicina: Revisão Sistemática da Literatura. Healthcare (Basel), v. 12, n. 18, p. 1916, 2024.
17. BILAL, Muhammad; HAMZA, Ameer; MALIK, Nadia. Natural Language Processing for Analyzing Electronic Health Records and Clinical Notes in Cancer Research: A Review. Preprint, Elsevier, 2024. Disponível em: https://arxiv.org/abs/2410.22180. Acesso em: 7 out. 2025.
18. ZHAO, Xueying; LI, Wei; WANG, Chen; ZHANG, Qi; LIU, Xinxin; JIANG, Rui; LI, Yong; ZHANG, Jian. A Multi-source and Multi-modal Data Fusion Framework for Predicting In-hospital Mortality of ICU Patients. Computers in Biology and Medicine, v. 155, 2023, p. 106643. doi: 10.1016/j.compbiomed.2023.106643