RevisãoZhili Niu, Xiandong Kuang, Juanjuan Chen, Xin Cai e Pingan Zhang*

A aplicação e desafios do ChatGPT na medicina laboratorial

Recebido em 1 de maio de 2025; aceito em 24 de julho de 2025; publicado online em 15 de agosto de 2025

Resumo:

Nos últimos anos, com o rápido desenvolvimento da tecnologia de inteligência artificial, os chatbots demonstraram um potencial significativo no campo médico, particularmente em laboratórios médicos. Este estudo analisa sistematicamente as vantagens e desafios dos chatbots neste campo e investiga suas potenciais aplicações no diagnóstico de doenças. No entanto, a confiabilidade e a natureza científica dos chatbots são influenciadas por vários fatores, incluindo qualidade dos dados, viés do modelo, proteção da privacidade e requisitos de feedback dos usuários. Para garantir a precisão e a confiabilidade do conteúdo gerado, é essencial não apenas confiar em estruturas legais, como o Ato de IA da UE, para a proteção necessária, mas também empregar duas ferramentas de avaliação, METRICS e CLEAR. Essas ferramentas são projetadas para avaliar de forma abrangente a qualidade das informações de saúde geradas por IA, proporcionando assim uma base teórica sólida e suporte para a prática clínica.

Palavras-chave: ChatGPT; chatbots; laboratório médico; vantagens; desafios

Introdução

A inteligência artificial (IA), primeiramente conceitualizada em 1950 pela pergunta seminal de Alan Turing, “As máquinas podem pensar?”, lançou as bases teóricas para este campo . O termo “inteligência artificial” foi oficialmente cunhado em 1956 por John McCarthy e colegas na Conferência de Dartmouth, marcando o nascimento da IA como uma disciplina de pesquisa independente . A IA abrange um amplo espectro, desde a inteligência artificial geral (AGI), que busca replicar a inteligência humana, até a inteligência artificial estreita (ANI), focada em tarefas específicas. Inicialmente, a IA dependia de algoritmos baseados em regras, como regras “se-então”, com esses sistemas não adaptativos ainda em uso em muitos ambientes clínicos hoje. No entanto, os avanços tecnológicos mudaram as aplicações de IA para métodos mais sofisticados, incluindo algoritmos de aprendizado de máquina (ML) e aprendizado profundo (DL), melhorando significativamente a análise de dados, o reconhecimento de imagens e a previsão de doenças em diagnósticos médicos. Com os laboratórios clínicos gradualmente alcançando automação, os resultados de testes massivos e complexos gerados diariamente superam a capacidade de processamento de um único cérebro humano, tornando a aplicação de IA cada vez mais importante. A IA pode analisar profundamente dados estruturados de alta qualidade em laboratórios clínicos, impulsionando a transformação dos métodos de diagnóstico . Modelos fundamentais de IA generativa podem simplificar fluxos de trabalho, melhorar a eficiência e reduzir os custos de saúde; aprimorar a medicina personalizada ao melhorar a precisão das

previsões de risco de doenças e resultados; e melhorar a alfabetização em saúde pública ao capacitar os pacientes com informações de saúde facilmente compreensíveis para melhor autogerenciamento . Nos últimos anos, o desenvolvimento de grandes modelos de linguagem (LLMs), como transformadores pré-treinados (GPT), BERT e o modelo de linguagem Pathways (PaLM), teve um impacto profundo em vários campos, incluindo geração de texto, tradução automática e design de conteúdo criativo. Os LLMs revolucionaram as interações entre pacientes e médicos ao simular conversas semelhantes às humanas, alterando significativamente os padrões de comunicação tradicionais na saúde . Embora os chatbots de IA tenham o potencial de melhorar a educação médica, abordar prontamente perguntas rotineiras sobre testes laboratoriais e ajudar na interpretação de resultados de testes, suas aplicações ainda estão em estágios iniciais. Os riscos e desafios associados exigem mais exploração . Os pacientes frequentemente buscam informações na internet quando explicações oportunas sobre os resultados de testes laboratoriais não são fornecidas, o que pode levar a desinformação e riscos à saúde . Além disso, o desempenho de diferentes modelos de chatbot varia significativamente, influenciado por fatores como configurações do modelo, diversidade de prompts e metodologias de avaliação .

Chatbots de IA Zhili Niu e Xiandong Kuang contribuíram igualmente para este trabalho.

*Autor correspondente: Pingan Zhang, Departamento de Laboratório Clínico, Instituto de Medicina Translacional, Hospital Renmin da Universidade de Wuhan, Wuhan, Província de Hubei, Hubei, R.P. China, E-mail: zhangpingan927@163.com

Zhili Niu, Xiandong Kuang, Juanjuan Chen e Xin Cai, Departamento de Laboratório Clínico, Instituto de Medicina Translacional, Hospital Renmin da Universidade de Wuhan, Wuhan, Hubei, R.P. China

Adv Lab Med 2025; aop

Acesso Aberto. © 2025 os autores, publicado por De Gruyter.

Este trabalho está licenciado sob a Licença Internacional Creative Commons Attribution 4.0.

AI chatbots pode gerar informações aparentemente credíveis, mas na verdade imprecisas, necessitando de uma compreensão abrangente de suas forças e limitações por parte de clínicos e pessoal de laboratório para utilizá-las e gerenciá-las efetivamente . O objetivo desta revisão é resumir as vantagens e limitações dos chatbots em laboratórios médicos e explorar os desafios e direções futuras de pesquisa em ambientes de laboratório clínico. Visa fornecer referências valiosas para a equipe de laboratório clínico que aplica chatbots, promovendo assim a aplicação adicional da tecnologia de inteligência artificial no campo médico.

Demandas e status atual da inteligência artificial

Relatórios de medicina laboratorial desempenham um papel crucial na tomada de decisões clínicas, influenciando significativamente diagnósticos e escolhas de tratamento, além de desempenhar um papel vital na gestão da saúde do paciente. No entanto, a complexidade desses relatórios muitas vezes confunde não especialistas, levando muitos pacientes a buscar explicações de chatbots ao enfrentar problemas de saúde. Pesquisas indicam que 78% dos usuários do ChatGPT tendem a utilizá-lo para autodiagnóstico, sublinhando a demanda significativa por fontes confiáveis de informações sobre saúde . De acordo com estudos de Giardina et al., 46% dos pacientes recorrem a buscas na web para entender seus resultados de testes, no entanto, esse comportamento também reflete os desafios potenciais que enfrentam na avaliação da gravidade desses resultados . A ausência de informações diagnósticas claras e oportunas pode exacerbar a ansiedade e o mal-entendido dos pacientes. Além disso, pesquisas de Kopanitsa demonstram que pacientes que recebem explicações geradas automaticamente são significativamente mais propensos a fazer acompanhamento (71%) em comparação com aqueles que apenas recebem resultados de testes (49%). Os pacientes geralmente apreciam a pontualidade dessas explicações, destacando como métodos de comunicação inovadores podem melhorar significativamente as experiências e os resultados de saúde dos pacientes . Características hierárquicas significativas são observadas na percepção e aceitação da IA entre a equipe clínica . A maioria dos entrevistados apoia a IA como uma ferramenta suplementar, valorizando particularmente seu papel na análise de dados. Profissionais jovens (<45 anos) exibem níveis de conhecimento sobre IA autoavaliados relativamente mais baixos em comparação com seus pares, no entanto, demonstram uma inclinação significativamente maior para o aprendizado contínuo. Por outro lado, a formação educacional desempenha um papel fundamental nos níveis de ansiedade, com pós-graduados expressando preocupações reduzidas sobre erros relacionados à IA devido a capacidades de compreensão técnica mais fortes. Notavelmente, a equipe de laboratório de linha de frente exibe uma preocupação profunda com a substituição no trabalho (32%), maior do que os 25% relatados por aqueles em funções gerenciais. Modelos de IA multimodal podem mitigar as limitações da alfabetização em IA entre a equipe de laboratório, integrando diversos tipos de dados (por exemplo, imagens de morfologia celular, biomarcadores celulares e parâmetros bioquímicos) em uma estrutura analítica unificada . Um sistema multimodal poderia combinar redes neurais convolucionais (CNNs) para extrair características espaciais de esfregaços de sangue, modelos baseados em grafos para analisar redes de interação celular e algoritmos de regressão para interpretar tendências bioquímicas . Essa integração reduz a dependência da síntese manual de dados, automatizando tarefas complexas, como identificar células displásicas na leucemia, enquanto correlaciona anomalias morfológicas com desequilíbrios bioquímicos (por exemplo, aumento da desidrogenase láctica). Para melhorar a acessibilidade, esses modelos podem ser incorporados em plataformas amigáveis ao usuário com interfaces intuitivas – como uploads de imagens arrastar e soltar e geração automática de relatórios – que abstraem complexidades técnicas. Além disso, a incorporação de componentes de IA explicável (XAI), como mapas de calor destacando características celulares críticas ou resumos em linguagem natural de correlações bioquímicas,

capacita o pessoal de laboratório a validar resultados sem exigir uma profunda expertise técnica. Ao simplificar fluxos de trabalho e fornecer insights contextualizados, modelos multimodais preenchem a lacuna entre as capacidades avançadas de IA e as operações práticas de laboratório, promovendo confiança e adoção mesmo entre usuários com treinamento limitado em IA. Atualmente, as reservas de conhecimento em IA dentro das equipes de laboratório são alarmantemente inadequadas, com apenas 10,8% das equipes demonstrando um domínio robusto de IA, e uma proporção notavelmente menor de cientistas de dados em hospitais terciários em comparação com instituições médicas privadas. Esse contraste acentuado é ainda mais destacado pelos 89,7% dos entrevistados que sublinham uma demanda urgente por treinamento em IA. No entanto, 47,2% dos laboratórios enfrentam desafios significativos na implementação efetiva de tecnologias de IA, principalmente devido à inadequação das equipes de suporte de TI . Discrepâncias entre as partes interessadas em relação à alocação de responsabilidade por erros médicos relacionados à IA são significativas . Os resultados da pesquisa revelam que 66,7% dos entrevistados defendem a responsabilidade compartilhada entre usuários e fabricantes, enquanto alguns argumentam que os fabricantes devem assumir a responsabilidade exclusiva por produtos que apresentem falhas de design. Os clínicos tendem a atribuir responsabilidade conjunta a fabricantes e instituições de saúde, afirmando que os fabricantes devem assumir a responsabilidade primária se as ferramentas de IA tiverem passado por validação rigorosa e cumprirem os padrões aplicáveis .

Integração de interfaces HL7-FHIR/LIS e considerações de privacidade-segurança

A implementação prática de chatbots de IA em ambientes laboratoriais requer uma integração robusta com a infraestrutura existente, particularmente através de interfaces padronizadas, como os Recursos de Interoperabilidade Rápida em Saúde do Nível 7 (HL7-FHIR) e Sistemas de Informação Laboratorial (LIS) . HL7-FHIR permite a troca de dados interoperáveis entre modelos de IA e bancos de dados laboratoriais, garantindo acesso em tempo real a resultados de testes estruturados, históricos de pacientes e critérios diagnósticos . A interpretação impulsionada por IA de perfis bioquímicos poderia ser aprimorada por meio de comunicação bidirecional com LIS, permitindo a sinalização automatizada de valores críticos ou análise contextual de dados longitudinais . No entanto, tal integração exige protocolos robustos de cibersegurança para proteger dados de saúde sensíveis. Os requisitos básicos de privacidade e segurança incluem criptografia de ponta a ponta das transmissões de dados, controles de acesso baseados em funções em conformidade com regulamentos como o Regulamento Geral sobre a Proteção de Dados (GDPR) e a Lei de Portabilidade e Responsabilidade de Seguros de Saúde (HIPAA), e auditorias regulares para detectar vulnerabilidades.

A aplicação de chatbots de inteligência artificial na medicina laboratorial

Atualmente, a IA alcançou uma ampla aplicação em áreas como radiologia para reconhecimento de imagens, no entanto, sua utilização na medicina laboratorial ainda está em estágios iniciais. Isso se deve principalmente à interpretação de relatórios laboratoriais que envolvem um vasto

número de variáveis quantitativas e qualitativas complexas, como sintomas, histórico médico e resultados de testes . Esses fatores impõem demandas mais altas sobre a complexidade e a precisão dos modelos de IA. Desde o lançamento do ChatGPT em 2022, ele tem atraído atenção significativa no campo médico. Estudos relevantes investigaram seu desempenho em exames de licenciamento médico, praticidade em abordar perguntas de pacientes e capacidade em auxiliar médicos na resolução de problemas clínicos, demonstrando assim o potencial de aplicação da IA na medicina laboratorial .

Capacidades do ChatGPT em abordar perguntas de laboratório clínico

Apesar de não ter passado por treinamento especializado em dados médicos, o ChatGPT demonstrou uma praticidade inicial no campo médico. Um estudo de Munoz-Zuluaga et al. envolveu a submissão de 65 perguntas cobrindo múltiplos tópicos ao ChatGPT para avaliar suas capacidades e precisão em responder perguntas relacionadas à medicina laboratorial . Os resultados mostraram que o ChatGPT respondeu corretamente 50,7% das perguntas, forneceu respostas incompletas ou parcialmente corretas para 23,1%, entregou informações incorretas ou enganosas para 16,9% e produziu respostas irrelevantes para 9,3%. Notavelmente, as respostas corretas estavam mais concentradas em perguntas relacionadas ao conhecimento médico ou técnico fundamental (59,1%), enquanto os erros eram mais comuns em perguntas referentes a procedimentos laboratoriais ou regulamentos (31%). Embora o GPT-4 mostre melhorias significativas em precisão, limitações em certos domínios permanecem evidentes. Um estudo de Girton et al. avaliou a capacidade do ChatGPT em abordar 49 perguntas clínicas de laboratório de pacientes reais em plataformas de mídia social, como Reddit e Quora, e comparou essas respostas com as fornecidas por estudantes de medicina . Os revisores tendiam a preferir as respostas do ChatGPT em relação às dos profissionais médicos. Isso destaca o progresso do ChatGPT em lidar com perguntas clínicas complexas, ao mesmo tempo em que revela lacunas na educação profissional médica nesta área. Aproximadamente metade das respostas dos estudantes de medicina foram classificadas como "ruins", em comparação com menos de 10% das respostas do ChatGPT recebendo classificações semelhantes . A ênfase limitada no conhecimento de laboratório clínico na educação médica pode contribuir para que até mesmo profissionais médicos experientes tenham dificuldades em fornecer respostas de alta qualidade a perguntas do mundo real. Outro estudo comparou respostas de três chatbots (ChatGPT, Gemini e Le Chat) com aquelas de médicos certificados em consultas online . No geral, os chatbots superaram os profissionais médicos online na interpretação de resultados laboratoriais. Enquanto os médicos online classificaram-se em primeiro lugar em 60% dos casos, o Gemini recebeu pontuações mais baixas em apenas 39% dos casos. O ChatGPT demonstrou qualidade e precisão comparáveis aos praticantes humanos. No entanto, os chatbots mostraram maiores tendências para superestimar a gravidade clínica (22 – 33% de incidência) em comparação com a taxa de superestimação de 1,0% dos médicos. Isso destaca os desafios dos chatbots em processar informações contextuais complexas e interpretar dados laboratoriais, com a falta de

padrões de referência unificados potencialmente levando a interpretações inconsistentes dos dados de testes dos pacientes.

Implementação de chatbots em diagnósticos laboratoriais médicos

Avanços recentes em inteligência artificial têm atraído interesse significativo de pesquisa em relação à interpretação de dados bioquímicos. Kaftan AN et al. demonstraram níveis de precisão variados entre modelos de IA (Copilot, Gemini) na análise de dados bioquímicos . O Copilot exibiu desempenho superior em todas as métricas de avaliação, alcançando vantagens estatisticamente significativas na interpretação de parâmetros bioquímicos em comparação com outros modelos. Isso sugere que as capacidades aprimoradas de processamento de dados do Copilot podem derivar de sua arquitetura algorítmica sofisticada. Enquanto outros modelos demonstraram desempenho subótimo em relação ao Copilot, suas competências específicas de domínio mantêm relevância prática para aplicações clínicas direcionadas . O domínio da microbiologia emergiu como um foco estratégico para a integração de inteligência artificial (IA) devido às suas exigências rigorosas por precisão diagnóstica e processamento rápido de dados. A IA demonstra potencial transformador em melhorar a eficiência da tomada de decisões clínicas e otimizar o manejo de doenças infecciosas, melhorando assim os resultados para os pacientes. No entanto, os sistemas de IA requerem um refinamento adicional na manipulação de cenários clínicos complexos e na geração de respostas contextualmente apropriadas para garantir relevância clínica . Notavelmente, o ChatGPT-4.0 apresenta limitações significativas ao abordar perguntas relacionadas a Testes de Suscetibilidade Antimicrobiana (AST) . O modelo recomendou inadequadamente clindamicina para infecções por Enterococcus e não conseguiu incorporar metodologias padrão para avaliação de suscetibilidade à polimixina. Tal orientação errônea poderia desviar a prática clínica, resultando em regimes terapêuticos ineficazes e potencial dano ao paciente. Li Y et al. conduziram uma avaliação sistemática de dois LLMs, ChatGPT e Google Gemini, ao abordar perguntas clínicas relacionadas ao HBV . O ChatGPT-4.0 demonstrou desempenho geral superior, alcançando 80,8% de precisão em perguntas baseadas em evidências em comparação com o Google Gemini (73,1%). A análise específica de domínio revelou a utilidade diagnóstica aprimorada do ChatGPT-4.0 na interpretação sorológica do HBV, enquanto o Google Gemini forneceu descrições mais abrangentes das manifestações clínicas. Todos os modelos exibiram limitações críticas na transmissão de estratégias de prevenção do HBV, particularmente em relação à atualidade das informações sobre vacinas. Notavelmente, apenas o Google Gemini fez referência ao consenso atual recomendando a vacinação neonatal contra a hepatite B dentro de 12 horas após o nascimento. Além disso, apesar das variações de precisão, tanto as respostas do ChatGPT quanto do Gemini excederam o nível de legibilidade recomendado para a 8ª série (Flesch-Kincaid grau 10,2 – 11,4), potencialmente comprometendo a compreensão do paciente. Esses achados ressaltam a necessidade de verificação por parte dos clínicos das saídas dos LLMs para garantir tanto a interpretabilidade quanto a validade clínica. Este estudo comparativo fornece insights cruciais sobre as forças e limitações específicas dos modelos para a implementação clínica baseada em evidências.

Desafios multidimensionais dos chatbots de saúde na prática clínica

Embora os chatbots sejam projetados e posicionados como ferramentas de conversação em vez de conselheiros médicos ou sistemas de suporte à decisão, eles demonstraram capacidades impressionantes na detecção e interpretação de anomalias laboratoriais. Essa habilidade é particularmente evidente em seu processamento e análise eficientes de vastos conjuntos de dados, especialmente com modelos avançados de linguagem como o ChatGPT. No entanto, essas potenciais capacidades são acompanhadas por limitações significativas, em parte porque esses modelos não foram especificamente treinados ou otimizados para relatórios de medicina laboratorial.

Desafios operacionais do ChatGPT na interpretação laboratorial clínica

A avaliação do Grupo de Trabalho da Federação Europeia de Química Clínica e Medicina Laboratorial (EFLM) revela limitações significativas na utilidade clínica do ChatGPT, apesar de sua capacidade de gerar análises de resultados laboratoriais "amplamente corretas e relevantes para a segurança" . Notavelmente, o ChatGPT falha em identificar marcadores de doenças subclínicas dentro de intervalos de referência normais – por exemplo, ele ignora que a elevação da GGT pode não indicar conclusivamente lesão hepática, e distribuições normais de subpopulações de leucócitos não garantem a integridade do sistema imunológico. Essa análise isolada de parâmetros corre o risco de perder indicadores patológicos precoces. Deficiências críticas emergem no reconhecimento de fatores pré-analíticos. Enquanto o ChatGPT sinaliza adequadamente o risco de diabetes com glicose elevada e HbA1c, ele negligencia variáveis cruciais de pré-coleta (por exemplo, estado de jejum) ao interpretar resultados discordantes (glicose elevada com HbA1c normal). Além disso, o ChatGPT demonstra síntese inadequada de

multi-analyte profiles – interpretações de marcadores de função hepática (ALT, AST, bilirrubina, GGT) carecem da integração do contexto clínico essencial necessária para um raciocínio diagnóstico abrangente . Além disso, o ChatGPT não consegue analisar minuciosamente os riscos à saúde associados aos resultados dos testes. Por exemplo, em casos de anemia severa ou anomalias no perfil lipídico, ele apenas aconselha os pacientes a consultar um médico, mas não os informa claramente sobre a gravidade clínica potencial dessas condições, que podem representar ameaças significativas à saúde dos pacientes. Por fim, o ChatGPT é incapaz de distinguir efetivamente entre valores atípicos e valores críticos de limiar clínico. Essa limitação pode levar a erros médicos em cenários clínicos críticos. A incapacidade de diferenciar entre valores anormais e valores clínicos de nível de emergência é particularmente preocupante, uma vez que tais distinções são cruciais para intervenções médicas oportunas e apropriadas.

Diversidade e variabilidade de chatbots No ambiente atual de ecossistemas de chatbots diversos, a variabilidade entre modelos e os desafios que eles apresentam em comparações entre estudos tornaram-se particularmente significativos. Modelos de IA generativa variam muito em arquitetura, configurações e objetivos, resultando em diferenças substanciais em seu

desempenho e qualidade ao gerar conteúdo. Seu design arquitetônico e capacidades influenciam diretamente a precisão e a eficácia de suas saídas. Notavelmente, existem diferenças significativas de desempenho entre diferentes modelos de IA generativa para tarefas específicas. Por exemplo, o Bing demonstrou a maior precisão e especificidade na previsão de interações medicamentosas, superando significativamente o Bard e o ChatGPT-4, destacando que certos modelos podem ter vantagens inerentes em cenários de aplicação específicos . Além disso, a variabilidade dentro do mesmo modelo não pode ser ignorada. O mesmo modelo pode produzir resultados inconsistentes sob diferentes configurações, dados de entrada ou estratégias de geração, especialmente no domínio da saúde pública, onde tal variabilidade pode ter implicações significativas para a tomada de decisões . Diferenças no desempenho do modelo não apenas afetam a eficácia da disseminação de informações, mas também influenciam diretamente a experiência e a satisfação do usuário. Por exemplo, descobriu-se que o Bard fornece as informações mais facilmente compreensíveis sobre rinoplastia, seguido pelo ChatGPT e pelo Bing . Portanto, pesquisadores e usuários devem exercer uma seleção cuidadosa ao aplicar modelos de IA generativa para garantir sua adequação a tarefas específicas e necessidades dos usuários.

Riscos associados a atualizações de versões de chatbot Embora atualizações contínuas de chatbots visem aprimorar suas capacidades e desempenho, elas também apresentam desafios em termos de consistência de dados, qualidade de dados e confiabilidade. Atualizações de modelos podem introduzir novos conhecimentos enquanto arriscam a referência a dados desatualizados, particularmente em campos que evoluem rapidamente, como a medicina. Por exemplo, os dados usados para treinar o ChatGPT não incluíram o valor de referência de chumbo no sangue de 3,5 μg/dL do CDC para crianças. Em contraste, o CopyAI foi capaz de fornecer valores de referência precisos de chumbo no sangue, indicando que diferenças nas atualizações de dados de treinamento entre chatbots podem levar a resultados de julgamento inconsistentes . Além disso, as atualizações nem sempre resultam em desempenho aprimorado, pois a precisão em certas tarefas pode diminuir. Um estudo realizado pela Universidade de Stanford e pela Universidade da Califórnia, Berkeley, descobriu que, embora o GPT-4.0 possa fornecer informações mais abrangentes em certas áreas, sua precisão em tarefas como matemática, codificação e raciocínio visual na verdade diminuiu . Esse fenômeno destaca a importância de avaliar continuamente o desempenho dos LLMs, particularmente em campos como a medicina laboratorial, onde a precisão é crítica.

Desafios da consistência na geração de conteúdo por modelos generativos de IA A consistência da saída de chatbots na geração de conteúdo merece atenção especial, particularmente em campos especializados como a medicina. Pesquisas indicam que pequenas diferenças na redação de prompts ou informações de fundo podem levar a variações significativas no conteúdo gerado, potencialmente afetando a confiabilidade em aplicações práticas . Um estudo de Kochanek e colegas demonstrou que a consistência das respostas do GPT-4.0 ao responder a mesma pergunta ao longo de quatro dias foi de 85 – 88%, respectivamente,

destacando a influência das incertezas na geração de conteúdo . No campo médico, onde alta consistência na saída é crítica, a variabilidade nas respostas pode representar riscos notáveis. Uma avaliação adicional da reprodutibilidade e precisão do ChatGPT na medicina laboratorial é essencial para garantir sua confiabilidade como referência em aplicações médicas . Além disso, modelos de IA generativa podem ser influenciados por preconceitos culturais e socioculturais inerentes aos seus dados de treinamento, levando a inconsistências no conteúdo gerado em diferentes contextos culturais ou linguísticos. Por exemplo, Wang e colegas descobriram que o ChatGPT teve um desempenho significativamente melhor em inglês do que em chinês durante o Exame de Licenciamento de Farmácia de Taiwan . Alfertshofer e colegas observaram ainda que o ChatGPT apresentou diferenças substanciais de desempenho em seis países em exames de licenciamento médico, destacando o impacto de fatores nacionais e linguísticos . Esses achados ressaltam a necessidade de uma investigação mais profunda e otimização de modelos de IA generativa em ambientes multilíngues e interculturais para melhorar sua adaptabilidade e precisão em cenários diversos.

Desafios na qualidade da disseminação de informações de saúde

Os chatbots estão desempenhando um papel cada vez mais importante na disseminação de informações de saúde, mas a clareza, concisão e relevância de suas respostas requerem atenção cuidadosa . A clareza exige que o conteúdo gerado seja fácil de entender, evitando o uso de jargão médico complexo para que o público em geral possa compreender os pontos principais sem esforço. Simultaneamente, a concisão requer comunicação direta sem extensão desnecessária, melhorando a adequação das informações de saúde e ajudando o público a melhorar sua literacia em saúde, capacitando-os a tomar decisões informadas sobre saúde. Além disso, a relevância do conteúdo gerado por IA é crucial. Informações precisas e relevantes ajudam a prevenir mal-entendidos, particularmente em cenários onde os pacientes não podem acessar explicações presenciais de médicos, evitando ansiedade desnecessária ou riscos à saúde causados por desinformação . Priorizar a relevância da informação previne a sobrecarga de informações, garantindo a entrega clara de mensagens de saúde críticas. Informações não relacionadas a consultas de saúde podem prejudicar a compreensão das pessoas sobre suas condições de saúde, aumentando sua confusão ao lidar com questões de saúde pessoal. No entanto, a disseminação de informações de saúde via IA deve equilibrar as necessidades diferentes de pacientes e profissionais de saúde. Os pacientes geralmente preferem respostas curtas e diretas para compreender rapidamente informações-chave, enquanto os profissionais de saúde tendem a favorecer respostas detalhadas e abrangentes. Essa divergência de necessidades cria uma espada de dois gumes: enquanto informações abundantes ajudam os profissionais em análises minuciosas, podem sobrecarregar usuários comuns, levando a dificuldades de compreensão e ansiedade. Independentemente do método de apresentação, garantir a completude das informações é primordial. A ausência de detalhes necessários pode levar a diagnósticos incorretos ou autodiagnósticos imprecisos, exacerbando riscos à saúde. Por exemplo, pacientes com literacia em saúde limitada podem ter dificuldade em interpretar resultados de testes

laboratoriais com precisão, como identificar níveis elevados de colesterol em painéis lipídicos, potencialmente levando a problemas de saúde graves, como doenças cardíacas ou derrames . Portanto, melhorar a capacidade dos pacientes de entender resultados de testes laboratoriais é de extrema importância.

Riscos de chatbots fornecendo informações de "alucinação"

A aplicação de chatbots na disseminação de informações médicas está se tornando cada vez mais comum, mas também levantou preocupações significativas em relação ao risco de desinformação. Embora ferramentas de IA generativa como o GPT ofereçam certas vantagens na fornecimento de informações, sua incapacidade de explicar processos de tomada de decisão prejudica sua capacidade de identificar e corrigir efetivamente preconceitos ou erros dentro do modelo. No campo médico, a desinformação gerada por IA pode levar a consequências graves, incluindo autodiagnósticos incorretos, atrasos no atendimento médico, a propagação de doenças potencialmente prejudiciais e uma perda de confiança do público em profissionais e instituições de saúde. Por exemplo, o GPT forneceu descrições imprecisas de dispositivos de teste de troponina de alta sensibilidade aprovados pelo FDA. Essas respostas de "alucinação" não são incidentes isolados, destacando as limitações dos modelos de IA generativa em garantir a precisão das informações . Portanto, garantir a precisão, confiabilidade e credibilidade das informações médicas geradas por IA é de suma importância. Os desenvolvedores devem priorizar esse objetivo durante o design e treinamento de modelos de IA, uma vez que estudos existentes demonstram que a geração de informações imprecisas é um fenômeno generalizado em tais ferramentas.

Questões legais e éticas nos dados de treinamento de chatbots

Questões legais e éticas em torno dos dados de treinamento de chatbots são preocupações críticas no desenvolvimento de modelos de inteligência artificial (IA) generativa . Essas questões não se referem apenas à origem dos dados, mas também envolvem seu uso e aceitabilidade, impactando diretamente a transparência e credibilidade do modelo. A transparência dos dados é de particular importância, garantindo que usuários e pesquisadores possam entender os métodos de coleta e utilização de dados, validando assim a confiabilidade e integridade científica do modelo. Além disso, preocupações éticas não devem ser negligenciadas, especialmente ao usar materiais protegidos por direitos autorais e dados clínicos, pois é imperativo garantir conformidade legal.

Aquisição e uso de dados enquanto salvaguarda os direitos de privacidade dos provedores de dados.

Modelos de IA generativa como o ChatGPT são treinados com dados baseados na internet, inevitavelmente herdando preconceitos presentes nos dados de treinamento e refletindo-os em suas saídas . Estudos mostraram que LLMs exibem preconceitos relacionados a gênero, raça e outros fatores sociais, o que não apenas compromete a equidade das saídas do modelo, mas

também exacerba desigualdades sociais. Esses preconceitos estão intimamente ligados a questões como lacunas de dados, tamanhos de amostra insuficientes e preconceitos inerentes em conjuntos de dados fundamentais . Se os dados de treinamento forem mais antigos, os preconceitos e erros potenciais podem ser amplificados nas saídas subsequentes do modelo . Esse fenômeno é particularmente evidente em registros existentes, especialmente em áreas-chave como design randomizado, destacando ainda mais os preconceitos potenciais nos processos de seleção de consultas.

Ato de IA da UE e diretrizes de suporte à decisão clínica da FDA

O Ato de Inteligência Artificial da União Europeia (AI Act), que entra em vigor em 12 de julho de 2024, estabelece o primeiro quadro regulatório abrangente do mundo para IA, priorizando o desenvolvimento ético, a proteção dos direitos fundamentais e a inovação centrada no ser humano . Adotando uma abordagem baseada em risco, o Ato categoriza os sistemas de IA em quatro níveis: risco inaceitável (proibido), alto risco (por exemplo, diagnósticos médicos, sujeitos a rigorosos requisitos de segurança, transparência e responsabilidade), IA de uso geral com risco sistêmico (obrigações de transparência) e baixo/não risco (regulação mínima) . Na saúde, padrões rigorosos para IA médica de alto risco enfatizam testes rigorosos, mitigação de preconceitos e monitoramento pós-mercado, embora lacunas persistam na abordagem dos direitos dos pacientes e da responsabilidade para aplicações de baixo risco. Os desafios incluem implementação desigual entre os estados membros, estruturas de responsabilidade ambíguas e limitações em evolução na detecção de conteúdo gerado por IA . O Ato também aborda as preocupações de direitos autorais da IA generativa, exigindo transparência nos dados de treinamento e salvaguardando a propriedade intelectual, enquanto isenta pesquisas pré-mercado . Espera-se que impactos extraterritoriais semelhantes ao GDPR remodelam as práticas globais de IA e necessitem de educação em ética de IA. Equilibrando regulação com inovação, o Ato de IA visa promover uma IA confiável, mas requer diretrizes mais claras, colaboração entre partes interessadas e padrões de saúde adaptados para garantir acesso equitativo e segurança do paciente. Atualmente, nenhum LLM é autorizado pela FDA como um dispositivo de suporte à decisão clínica. Pesquisas avaliaram se LLMs podem aderir a regulamentos no contexto de emergências clínicas. Os resultados revelaram que, embora LLMs fornecessem recomendações apropriadas de cuidados preventivos, 100% das respostas do GPT-4 e 52% do Llama-3 não estavam em conformidade em cenários críticos de tempo, oferecendo suporte semelhante a dispositivos ao sugerir diagnósticos e tratamentos específicos . Apesar de prompts baseados nas diretrizes da FDA, a conformidade não foi alcançada de forma consistente. Esses resultados destacam a necessidade urgente de novos quadros regulatórios para abordar os desafios únicos impostos pela IA generativa na saúde, uma vez que as diretrizes atuais são insuficientes para prevenir saídas não autorizadas semelhantes a dispositivos de LLMs.

Estabelecendo um quadro de avaliação abrangente para abordar os desafios dos chatbots de IA na saúde

Em resposta aos riscos impostos pelos chatbots de IA no campo médico, é imperativo desenvolver um quadro de avaliação abrangente. O primeiro passo para garantir a qualidade das saídas de IA envolve estabelecer critérios de avaliação rigorosos que abrangem a precisão, confiabilidade e relevância das informações. Além disso, testes de desempenho regulares ajudam a identificar falhas no sistema e garantem que os modelos de IA permaneçam alinhados com as atualizações no conhecimento médico. Mecanismos de feedback, que coletam informações de usuários e profissionais, fornecem dados valiosos para melhoria contínua. A inclusão de mecanismos de revisão profissional é igualmente crucial. O envolvimento de especialistas clínicos aprimora ainda mais a confiabilidade e a integridade científica dos sistemas de IA, garantindo que as informações geradas estejam alinhadas com a ética médica e os padrões de prática. Para alcançar esse objetivo, várias ferramentas e diretrizes foram desenvolvidas para auxiliar na avaliação e melhoria da qualidade das informações de saúde . No entanto, as ferramentas existentes de avaliação de informações de saúde não são especificamente adaptadas para avaliar a qualidade do conteúdo relacionado à saúde gerado por modelos de IA. Garantir a confiabilidade e a precisão da IA em ambientes laboratoriais médicos requer o uso de ferramentas de avaliação padronizadas. Atualmente, as ferramentas "METRICS" e "CLEAR" são amplamente reconhecidas como meios eficazes de avaliar informações de saúde geradas por IA . A lista de verificação METRICS fornece um quadro para projetar e relatar pesquisas padronizadas em IA, cobrindo nove temas principais: Modelo, Avaliação, Tempo, Alcance/Randômização, Fatores Individuais, Contagem e Especificidade de Prompts e Linguagem. Esses temas visam analisar de forma abrangente o design e o status operacional da IA.

Tabela 1: Entradas para MÉTRICAS.

Tipo de ferramenta	Conteúdo do item	Limitações	Classificação de escala
MÉTRICAS inclui nove itens: 1) Modelo 2) Avaliação 3) Tempo 4) Transparência 5) Amplitude 6) Randomização 7) Fatores individuais 8) Contagem 9) Especificidade dos prompts e linguagem	(1) Modelo: A(s) ferramenta(s) generativa(s) baseada(s) em IA usada(s) no registro incluído deve(m) mencionar explicitamente as configura(Salvo neste PC ra cada ferramenta. (2) Avaliação: A abordagem para avaliar a qualidade do conteúdo. Equilibra objetividade (resultados imparciais) com subjetividade. (3) Tempo: O tempo exato e a duração do teste do modelo de IA generativo. (4) Transparência: Transparência das fontes de dados (incluindo permissões para conteúdo protegido por direitos autorais). (5) Amplitude: O escopo dos tópicos testados (tópicos únicos/múltiplos relacionados/não relacionados; amplitude de consultas intra e intertópicos). (6) Randomização: Grau de randomização na seleção de tópicos para mitigar viés. (7) Individual: Papel subjetivo na avaliação do conteúdo e confiabilidade entre avaliadores (concordância/discordância). (8) Contagem: Número de consultas executadas por modelo (tamanho da amostra). (9) Especificidade: Indicadores: Fraseamento exato, mecanismos de feedback e ciclos de aprendizagem. Idioma: O(s) idioma(s) usado(s) nos testes, incluindo considerações culturais (por exemplo, adequação linguística e cultural).	(1) Escopo limitado da literatura: A literatura foi recuperada usando apenas uma única palavra-chave, o que pode resultar em omissões. (2) Viés de banco de dados e idioma: A literatura em inglês de bancos de dados como Scopus, PubMed e Google Acadêmico pode introduzir potencial viés de seleção. (3) Tópicos insuficientes: Devido às limitações da equipe de autores (formação unidisciplinar), tópicos transdisciplinares relevantes podem ter sido negligenciados. (4) Risco de subjetividade na avaliação: A seleção de tópicos e a avaliação dos autores (pontuação MÉTRICAS) podem envolver julgamentos subjetivos. (5) Igual ponderação na pontuação: Todas as MÉTRICAS adotaram o mesmo peso, desconsiderando as diferenças de importância entre os vários indicadores. Limitações de cobertura dos modelos: O foco concentrou-se exclusivamente em ChatGPT, Bing e Bard, sem considerar outros modelos generativos (como Claude).	Os itens do MÉTRICAS são pontuados usando uma escala Likert de 5 pontos da seguinte forma: 5=excelente, 4=muito bom, 3=bom, 2=satisfatório e 1=subótimo.

Tabela 2: Entradas para CLEAR.

Tipo de ferramenta	Conteúdo do item	Limitações	Classificação de escala
O CLEAR inclui cinco itens: 1) Completude do conteúdo 2) Ausência de informações falsas no conteúdo 3) Evidências que apoiam o conteúdo 4) Adequação do conteúdo(5) Relevância	(1) O conteúdo é suficiente? Completude significa que a informação é gerada de forma otimizada, nem excessiva nem deficiente. (2) O conteúdo é preciso? A geração de informações incorretas sobre saúde pelas ferramentas de IA pode levar a consequências negativas graves. (3) O conteúdo é baseado em evidências? Isso significa que as informações de saúde fornecidas pelo modelo de IA devem ser baseadas nos mais recentes avanços científicos e evitar viés, desinformação ou informações falsas. (4) O conteúdo é claro, conciso e fácil de entender? O conteúdo também deve fornecer uma explicação única e clara e ser bem organizado em uma sequência lógica para facilitar a compreensão. (5) O conteúdo está livre de informações irrelevantes? Relevância refere-se à necessidade de fornecer conteúdo de saúde preciso e pertinente.	1) Limitações da amostra: A avaliação da ferramenta CLEAR depende exclusivamente de um pequeno número de profissionais de saúde homogêneos, o que introduz viés; os dados do teste constituem declarações geradas artificialmente, sem validação contextual. (2) Validação insuficiente da ferramenta: A ferramenta CLEAR não foi comparada com outras ferramentas de avaliação de informações, como o DISC, para esclarecer seus pontos fortes e fracos; a eficácia da ferramenta deve ser confirmada por meio de testes em tópicos de saúde mais amplos, especialmente controversos. (3) Impacto da dinâmica do modelo de IA: Os modelos de IA são continuamente atualizados e iterados (por exemplo, mudanças do GPT-4 para o GPT-5 podem produzir resultados diferentes com a mesma entrada ao longo do tempo); o desempenho do modelo mostra-se sensível ao design dos prompts, e variações nos prompts podem levar a discrepâncias nos resultados. (4) Aplicabilidade externa limitada: As conclusões atuais baseiam-se em avaliações de um contexto disciplinar limitado, indicando a necessidade de validação que envolva especialistas multidisciplinares (como desenvolvedores de IA e representantes de pacientes).	Os itens do CLEAR são pontuados usando uma escala Likert de 5 pontos da seguinte forma: 5=excelente, 4=muito bom, 3=bom, 2=satisfatório e 1=abaixo do ideal.

Modelos, garantindo o rigor científico e a robustez de sua produção. Em contraste, a ferramenta CLEAR foca na avaliação sistemática de múltiplas dimensões do conteúdo gerado por IA, incluindo cinco critérios-chave: Completude do Conteúdo, Ausência de Informação Falsa no Conteúdo, Evidência que Apoia o Conteúdo, Apropriabilidade do Conteúdo e Relevância. Os significados específicos de cada critério estão detalhados na Tabela 1 e na Tabela 2, fornecendo aos avaliadores um guia claro para garantir que o conteúdo gerado por IA atenda a padrões de qualidade predefinidos.

Em um estudo de Malik Sallam et al., três modelos de IA generativa (ChatGPT, Microsoft Bing e Google Bard) foram avaliados usando a ferramenta METRICS . A média geral da pontuação METRICS entre esses modelos foi de 3,0 (DP 0,58). Analisando critérios individuais, o critério "Modelo" recebeu a maior pontuação média, seguido por "Especificidade". Por outro lado, as menores pontuações foram observadas para o critério "Randomização" (classificado como subótimo) e o critério "Fatores Individuais" (classificado como satisfatório). Separadamente, os cinco modelos de IA foram avaliados usando a ferramenta CLEAR em cinco tópicos distintos . O Microsoft Bing alcançou a maior pontuação média CLEAR (média: 24,4 ± 0,42), seguido pelo ChatGPT-4 (média: 23,6 ± 0,96), Google Bard (média: 21,2 ± 1,79) e, finalmente, ChatGPT-3.5 (média: 20,6 ± 5,20). Embora as ferramentas METRICS e CLEAR tenham contribuído para a avaliação do conteúdo gerado por IA, elas também têm certas limitações (como detalhado na Tabela 1 e na Tabela 2). Por exemplo, essas ferramentas podem não cobrir todas as variáveis em situações específicas ou apresentar aplicabilidade limitada em certos domínios especializados. Portanto, essas limitações requerem validação por meio de pesquisas adicionais para confirmar sua eficácia e adequação em aplicações práticas. Ao obter uma compreensão mais profunda e

abordar essas limitações, podemos aumentar a credibilidade das informações de saúde geradas por IA, proporcionando uma base sólida para a tomada de decisões médicas.

Conclusões e perspectivas

Embora os chatbots de IA enfrentem limitações, como a necessidade de conhecimento avançado, habilidades de raciocínio insuficientes e capacidade limitada para análise aprofundada das informações dos pacientes, seu potencial no diagnóstico de doenças e na medicina laboratorial não deve ser subestimado, como ilustrado na Figura 1, que fornece um resumo conciso das oportunidades e desafios associados aos chatbots. No campo da medicina laboratorial, chatbots de IA como o GPT são não apenas capazes de fornecer informações precisas, mas também podem alcançar ou superar o nível de resposta de profissionais médicos em certas áreas-chave .

Oportunidades e Desafios

Desafios

1. Qualidade dos Dados e Viés do Modelo

- Dados de treinamento desatualizados.
- Viéses herdados de conjuntos de dados históricos

2. Confiabilidade e Riscos Clínicos

- Análise isolada de biomarcadores sem contexto clínico
- Interpretação integrativa multiparâmetro limitada

3. Preocupações Éticas/Legais e de Privacidade

- Estruturas de responsabilização ambíguas
- Riscos de exposição a dados de saúde sensíveis

4. Variabilidade do Modelo e Qualidade da Informação

- Inconsistências de desempenho entre modelos de IA
- Variabilidade de resposta dentro de modelos únicos

5. Lacunas de Implementação e Treinamento

- Experiência insuficiente em IA nos laboratórios
- Infraestrutura de TI inadequada dificultando a implantação

Oportunidades

1. Processamento de Dados e Aprimoramento da Eficiência

- Análise automatizada de dados de testes massivos além dos limites cognitivos humanos
- Fluxos de trabalho otimizados para reduzir custos de saúde

2. Suporte Diagnóstico e Medicina Personalizada

- Precisão aprimorada da previsão de risco de doenças para tratamentos personalizados
- Interpretação de relatórios laboratoriais complexos com tecnologia de IA para reduzir a ansiedade do paciente

3. Educação do Paciente e Acessibilidade à Saúde

- Entrega simplificada de informações de saúde para melhor autogestão
- Explicações automatizadas de resultados de testes para melhorar as taxas de acompanhamento

4. Educação Médica e Assistência à Pesquisa

- Preenchendo lacunas de conhecimento em currículos médicos desatualizados
- Revisão acelerada de literatura e análise de dados (por exemplo, avaliação de LLMs)

5. Saúde Pública e Otimização de Recursos

- Capacidades aprimoradas de diagnóstico de atenção primária
- Redução das disparidades regionais de saúde

Processo intensivo em tempo com atualizações lentas, tornando desafiador para os estudantes compreenderem todos os pontos críticos. Portanto, os chatbots de IA têm o potencial de servir como ferramentas auxiliares importantes na educação médica, ajudando a preencher a lacuna na aquisição de conhecimento dos estudantes. No entanto, a resolução criativa de problemas continua a ser uma força única dos seres humanos. Para aproveitar totalmente o potencial da IA na medicina, a colaboração entre desenvolvedores de IA e profissionais de saúde é crucial. A medicina laboratorial é um campo em rápida evolução, com novas tecnologias e métodos de teste

sendo continuamente desenvolvidos e integrados à prática clínica. Se as limitações dos chatbots de IA puderem ser efetivamente abordadas, eles poderiam se tornar ferramentas indispensáveis para clínicos e técnicos de laboratório em seu trabalho diário. Superar os desafios existentes poderia levar a uma precisão diagnóstica aprimorada, qualidade de atendimento ao paciente melhorada e à promoção da medicina personalizada. A colaboração entre IA e especialistas médicos humanos será a base para o futuro desenvolvimento da medicina laboratorial.

Ética de pesquisa: Não aplicável.

Consentimento informado: O consentimento informado foi obtido de todos os indivíduos incluídos neste estudo.

Contribuições dos autores: Todos os autores aceitaram a responsabilidade pelo conteúdo total deste manuscrito e aprovaram sua submissão.

Uso de Modelos de Linguagem Grande, Ferramentas de IA e Aprendizado de Máquina: Para melhorar a linguagem.

Conflito de interesse: Os autores afirmam não ter conflito de interesse.

Financiamento da pesquisa: Nenhum declarado.

Disponibilidade de dados: Não aplicável.

Referências

1. Kaul V, Enslin S, Gross SA. História da inteligência artificial na medicina. Gastrointest Endosc 2020;92:807 – 12.2. Ferraro S, Panteghini M. O papel do laboratório em garantir solicitações de testes apropriadas. Clin Biochem 2017;50:555 – 61.3. Lavoie-Gagne O, Woo JJ, Williams RJ 3rd, Nwachukwu BU, Kunze KN, Ramkumar PN. Inteligência artificial como uma ferramenta para mitigar a carga administrativa, otimizar a cobrança, reduzir despesas relacionadas a seguros e credenciamento, e melhorar a garantia de qualidade dentro dos sistemas de saúde. Arthroscopy 2025;41:3270 – 5.4. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. Um framework de avaliação para o uso clínico de grandes modelos de linguagem em tarefas de interação com pacientes. Nat Med 2025;31:77 – 86.5. Giardina TD, Baldwin J, Nystrom DT, Sittig DF, Singh H. Percepções dos pacientes sobre o recebimento de resultados de testes via portais online: um estudo de métodos mistos. J Am Med Inform Assoc 2018;25:440 – 6.

1. Bar-Lev S, Beimel D. Números, gráficos e palavras, realmente entendemos os resultados dos testes laboratoriais acessíveis através dos portais dos pacientes? Isr J Health Pol Res 2020;9:58.

2. Chu SKW, Huang H, Wong WNM, Ginneken WFV, Hung MY. Qualidade e clareza das informações de saúde em sites de perguntas e respostas. Libr Inf Sci Res 2018;40.

3. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Qualidade das respostas de modelos de linguagem generativa de grande porte versus usuários pares para interpretar resultados de testes laboratoriais para pacientes leigos: estudo de avaliação. J Med Internet Res 2024;26:e56655.

4. O ChatGPT transformará a saúde? Nat Med 2023;29:505 – 6.

5. Eysenbach G. O papel do ChatGPT, modelos de linguagem generativa e inteligência artificial na educação médica: uma conversa com o ChatGPT e um chamado para submissões. JMIR Med Educ 2023;9:e46885.

6. Meskó B. Engenharia de prompt como uma habilidade emergente importante para profissionais médicos: tutorial. J Med Internet Res 2023;25:e50638.

7. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs google para consultas relacionadas à demência e outros declínios cognitivos: comparação de resultados. J Med Internet Res 2023;25:e48966.

8. Sallam M. Utilidade do ChatGPT na educação, pesquisa e prática em saúde: revisão sistemática sobre as perspectivas promissoras e preocupações válidas. Healthcare 2023;11. .

9. Rajpurkar P, Chen E, Banerjee O, Topol EJ. IA em saúde e medicina. Nat Med 2022;28:31 – 8.

10. Shahsavar Y, Choudhury A. Intenções dos usuários de usar o ChatGPT para autodiagnóstico e propósitos relacionados à saúde: estudo de pesquisa transversal. JMIR Hum Factors 2023;10:e47564.

11. Kopanitsa G. Estudo da atitude dos pacientes em relação à interpretação automática dos resultados de testes laboratoriais e sua influência na taxa de acompanhamento. BMC Med Inform Decis Mak 2022;22:79.

12. Sarwar S, Dent A, Faust K, Richer M, Djuric U, Van Ommeren R, et al. Perspectivas dos médicos sobre a integração da inteligência artificial na patologia diagnóstica. NPJ Digit Med 2019;2:28.

13. Ardon O, Schmidt RL. Atitudes dos funcionários de laboratórios clínicos em relação à inteligência artificial. Lab Med 2020;51:649 – 54.

14. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Confiança dos médicos na inteligência artificial: uma pesquisa online móvel. J Med Internet Res 2019;21:e12422.

15. Tarmissi K, Alsamri J, Maashi M, Asiri MM, Yahya AE, Alkharashi A, et al. Representações multimodais de aprendizado por transferência com algoritmo de otimização de cobra na classificação de células da medula óssea usando imagens histopatológicas biomédicas. Sci Rep 2025;15:14309.

16. Matek C, Krappe S, Münzenmayer C, Haferlach T, Marr C. Diferenciação altamente precisa das morfologias das células da medula óssea usando redes neurais profundas em um grande conjunto de dados de imagens. Blood 2021;138:1917 – 27.

17. Durant TJS, Olson EM, Schulz WL, Torres R. Redes neurais convolucionais muito profundas para classificação morfológica de eritrócitos. Clin Chem 2017;63:1847 – 55.

18. Cadamuro J, Carobene A, Cabitza F, Debeljak Z, De Bruyne S, van Doorn W, et al. Uma pesquisa abrangente sobre a adoção de inteligência artificial na medicina laboratorial europeia: utilização atual e perspectivas. Clin Chem Lab Med 2025;63:692 – 703.

19. Yu S, Jeon BR, Liu C, Kim D, Park HI, Park HD, et al. Preparação laboratorial para medicina digital na saúde 4.0: uma investigação sobre a conscientização e aplicações de big data e inteligência artificial. Ann Lab Med 2024;44:562 – 71.

20. Khullar D, Casalino LP, Qian Y, Lu Y, Chang E, Aneja S. Visões públicas vs visões de médicos sobre responsabilidade pela inteligência artificial na saúde. J Am Med Inform Assoc 2021;28:1574 – 7.

21. Abràmoff MD, Tobey D, Char DS. Lições aprendidas sobre IA autônoma: encontrando um caminho seguro, eficaz e ético através do processo de desenvolvimento. Am J Ophthalmol 2020;214:134 – 42.

22. Bazoukis G, Hall J, Loscalzo J, Antman EM, Fuster V, Armoundas AA. A inclusão da inteligência aumentada na medicina: uma estrutura para implementação bem-sucedida. Cell Rep Med 2022;3:100485.

23. Wen Y, Choo VY, Eil JH, Thun S, Pinto Dos Santos D, Kast J, et al. Troca de dados de composição corporal avaliados por tomografia computadorizada quantitativa usando recursos de interoperabilidade em saúde rápida como um passo necessário para a integração interoperável da triagem oportunista na prática clínica: estudo de desenvolvimento metodológico. J Med Internet Res 2025;27:e68750.

24. Dolin RH, Heale BSE, Alterovitz G, Gupta R, Aronson J, Boxwala A, et al. Introduzindo operações de genômica HL7 FHIR: uma abordagem amigável para desenvolvedores para integração de genômica-EHR. J Am Med Inform Assoc 2023;30:485 – 93.

25. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Recursos de interoperabilidade em saúde rápida (FHIR) para interoperabilidade em pesquisa em saúde: revisão sistemática. JMIR Med Inform 2022;10:e35724.

26. Tran DM, Thanh Dung N, Minh Duc C, Ngoc Hon H, Minh Khoi L, Phuc Hau N, et al. Status da adoção de tecnologia de saúde digital em 5 hospitais vietnamitas: avaliação transversal. JMIR Form Res 2025;9:e53483.

27. Burnside ES, Grist TM, Lasarev MR, Garrett JW, Morris EA. Inteligência artificial em radiologia: uma pesquisa de liderança. J Am Coll Radiol 2025;22:577 – 85.

28. Takagi S, Watari T, Erabi A, Sakaguchi K. Desempenho do GPT-3.5 e GPT-4 no exame de licenciamento médico japonês: estudo comparativo. JMIR Med Educ 2023;9:e48002.

29. Wang H, Wu W, Dou Z, He L, Yang L. Desempenho e exploração do ChatGPT em exames médicos, registros e educação em chinês: abrindo caminho para a IA médica. Int J Med Inf 2023;177:105173.

30. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Avaliação do desempenho de grandes modelos de linguagem para cuidados com a miopia: uma análise comparativa do ChatGPT-3.5, ChatGPT-4.0 e google bard. EBioMedicine 2023;95:104770.

31. Munoz-Zuluaga C, Zhao Z, Wang F, Greenblatt MB, Yang HS. Avaliando a precisão e a utilidade clínica do ChatGPT na medicina laboratorial. Clin Chem 2023;69:939 – 40.

32. Girton MR, Greene DN, Messerlian G, Keren DF, Yu M. ChatGPT vs profissional médico: analisando respostas a perguntas de medicina laboratorial nas redes sociais. Clin Chem 2024;70:1122 – 39.

33. Meyer A, Soleman A, Riese J, Streichert T. Comparação do ChatGPT, Gemini e Le Chat com interpretações de médicos sobre perguntas de laboratório médico de um fórum de saúde online. Clin Chem Lab Med 2024;62:2425 – 34.

34. Kaftan AN, Hussain MK, Naser FH. Precisão das respostas do ChatGPT 3.5 copilot e gemini na interpretação de dados laboratoriais bioquímicos: um estudo piloto. Sci Rep 2024;14:8233.

35. Sallam M, Al-Salahat K, Al-Ajlouni E. Desempenho do ChatGPT em cenários de casos orientados a laboratórios de microbiologia clínica diagnóstica. Cureus 2023;15:e50629.

36. Carey RB, Bhattacharyya S, Kehl SC, Matukas LM, Pentella MA, Salfinger M, et al. Orientações práticas para laboratórios de microbiologia clínica: implementando um sistema de gestão da qualidade no laboratório de microbiologia médica. Clin Microbiol Rev 2018;31. .

37. Genzen JR, Tormey CA. Consulta de patologia sobre a notificação de valores críticos. Am J Clin Pathol 2011;135:505 – 13.

38. Li Y, Huang CK, Hu Y, Zhou XD, He C, Zhong JW. Explorando o desempenho de grandes modelos de linguagem em perguntas relacionadas à infecção por hepatite B: um estudo comparativo. World J Gastroenterol 2025;31:101092.

39. Cadamuro J, Cabitza F, Debeljak Z, De Bruyne S, Frans G, Perez SM, et al. Potenciais e armadilhas do ChatGPT e modelos de inteligência artificial de linguagem natural para a compreensão dos resultados de testes de medicina laboratorial. Uma avaliação pelo grupo de trabalho sobre inteligência artificial (WG-AI) da Federação Europeia de Química Clínica e Medicina Laboratorial (EFLM). Clin Chem Lab Med 2023;61:1158 – 66.

40. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Avaliando a sensibilidade, especificidade e precisão do ChatGPT-3.5, ChatGPT-4, bing AI e bard em relação a ferramentas clínicas convencionais de interações medicamentosas. Drug Healthc Patient Saf 2023;15:137 – 47.

41. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Explorando o possível uso de chatbots de IA na educação em saúde pública: estudo de viabilidade. JMIR Med Educ 2023;9:e51421.

42. Seth I, Lim B, Xie Y, Cevik J, Rozen WM, Ross RJ, et al. Comparando a eficácia de grandes modelos de linguagem ChatGPT, BARD e bing AI em fornecer informações sobre rinoplastia: um estudo observacional. Aesthet Surg J Open Forum 2023;5:ojad084.

43. Abusoglu S, Serdar M, Unlu A, Abusoglu G. Comparação de três chatbots como assistentes para resolução de problemas em laboratório clínico. Clin Chem Lab Med 2024;62:1362 – 6.

44. Lingjiao Chen MZ, James Z. <Como o comportamento do chatgpt está mudando ao longo do tempo_.pdf>. Harv Data Sci Rev 2023;2024:2.

45. Giray L. Engenharia de prompt com ChatGPT: um guia para escritores acadêmicos. Ann Biomed Eng 2023;51:2629 – 33.

46. Khlaif ZN, Mousa A, Hattab MK, Itmazi J, Hassan AA, Sanmugam M, et al. O potencial e as preocupações do uso de IA na pesquisa científica: avaliação de desempenho do ChatGPT. JMIR Med Educ 2023;9:e47049.

47. Kochanek K, Skarzynski H, Jedrzejczak WW. Precisão e repetibilidade do ChatGPT com base em um conjunto de perguntas de múltipla escolha sobre testes objetivos de audição. Cureus 2024;16:e59857.

48. Wang YM, Shen HW, Chen TJ. Desempenho do ChatGPT no exame de licenciamento de farmacêuticos em Taiwan. J Chin Med Assoc 2023;86:653 – 8.

49. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Navegando pelos sete mares: uma comparação multinacional do desempenho do chatgpt em exames de licenciamento médico. Ann Biomed Eng 2024;52:1542 – 5.

50. Zikmund-Fisher BJ, Scherer AM, Witteman HO, Solomon JB, Exe NL, Fagerlin A. Efeito de âncoras de dano em exibições visuais de resultados de testes sobre percepções de urgência dos pacientes em relação a valores quase normais: estudo experimental. J Med Internet Res 2018;20:e98.

51. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: cinco prioridades para pesquisa. Nature 2023;614:224 – 6.

52. Sanderson K. GPT-4 chegou: o que os cientistas pensam. Nature 2023;615:773.

53. Murdoch B. Privacidade e inteligência artificial: desafios para proteger informações de saúde em uma nova era. BMC Med Ethics 2021;22:122.

54. Mijwil MM, Aljanabi M, Ali AH. ChatGPT: explorando o papel da cibersegurança na proteção de informações médicas. Mesopotamian J CyberSecurity 2023;2023. .

55. Chen Z. Ética e discriminação em práticas de recrutamento habilitadas por inteligência artificial. Palgrave Commun 2023;9:12.

56. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Considerações éticas sobre o uso do ChatGPT na saúde. J Med Internet Res 2023;25:e48009.

57. Liang P, Wu C, Morency LP, Salakhutdinov R. Rumo a entender e mitigar preconceitos sociais em modelos de linguagem. PMLR 2021;139:2640 – 3498.

58. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potenciais preconceitos em algoritmos de aprendizado de máquina usando dados de registros eletrônicos de saúde. JAMA Intern Med 2018;178:1544 – 7.

59. Tianyu Gao HY, Jiatong Y, Danqi C. <Habilitando grandes modelos de linguagem a gerar texto com citações.pdf>. ArXiv 2023;2023.

60. Gasser U. Um marco da UE para a governança da IA. Science 2023;380:1203.

61. März M, Himmelbauer M, Boldt K, Oksche A. Aspectos legais da inteligência artificial generativa e grandes modelos de linguagem em exames e teses. GMS J Med Educ 2024;41:Doc47.

62. Alvarado A. Lições da lei de IA da UE. Patterns (N Y) 2025;6:101183.

63. Weissman G, Mankowitz T, Kanter G. Não conformidade de grandes modelos de linguagem com as diretrizes da FDA para dispositivos de suporte à decisão clínica. Res Sq 2024. .

64. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: um instrumento para julgar a qualidade da informação de saúde escrita para o consumidor sobre opções de tratamento. J Epidemiol Community Health 1999;53:105 – 11.

65. Baur C, Prue C. O índice de comunicação clara do CDC é uma nova ferramenta baseada em evidências para preparar e revisar informações de saúde. Health Promot Pract 2014;15:629 – 37.

66. DeWalt DA, Broucksou KA, Hawk V, Brach C, Hink A, Rudd R, et al. Desenvolvendo e testando o kit de ferramentas de precauções universais de alfabetização em saúde. Nurs Outlook 2011;59:85 – 94.

67. Sallam M, Barakat M, Sallam M. Uma lista de verificação preliminar (METRICS) para padronizar o design e a apresentação de estudos sobre modelos baseados em inteligência artificial generativa na educação e prática em saúde: estudo de desenvolvimento envolvendo uma revisão da literatura. Interact J Med Res 2024;13:e54704.

68. Sallam M, Barakat M, Sallam M. Teste piloto de uma ferramenta para padronizar a avaliação da qualidade da informação de saúde gerada por modelos baseados em inteligência artificial. Cureus 2023;15:e49373.

Chatbots são softwares com inteligência artificial que simulam conversas humanas por texto ou voz para realizar tarefas como suporte ao cliente, agendamentos e consultas, podendo ser usados em sites, aplicativos e redes sociais

A "**IA da UE**" refere-se à Lei de Inteligência Artificial da UE, a primeira lei abrangente do mundo que estabelece regras para a IA, focando em um sistema baseado em risco para garantir a segurança e os direitos fundamentais. A legislação classifica os sistemas de IA por níveis de risco, proibindo práticas inaceitáveis, impondo requisitos mais rigorosos para sistemas de alto risco e estabelecendo regras para modelos de IA de propósito geral.