INSTITUTO DE MEDICINA INTEGRAL PROFESSOR FERNANDO FIGUEIRA - IMIP

PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇAO CIENTÍFICA A

PIBI-CNPq-IMIP - 2025/2026

INTELIGÊNCIA ARTIFICIAL COMO AVALIADORA DE HABILIDADES DE COMUNICAÇÃO CLÍNICA: UMA REVISÃO NARRATIVA ESTRUTURADA

Artigo cientifico submetido à XVI Congresso Estudantil da Faculdade Pernambucana de Saúde – FPS, como finalização do Programa de Iniciação Cientifica - PIBIC no ano de 2025/26 e como requisito parcial à apresentação do Trabalho de Conclusão de Curso.

Estudante bolsista: Rayne Lyanna Oliveira Dantas

Orientadora: Dra. Mirella Rebello Bezerra

Co-orientadora: Rafaella Christine Tenório de Arruda Macêdo

Estudantes colaboradores: Beatriz Bronzeado Clericuzi, João Victor Melo Soares, Lia Sancho Monteiro e Sabrina Vitória das Mercês Cavalcanti

Recife

2026

PARTICIPANTES DA PESQUISA

ORIENTADORA

Mirella Rebello Bezerra

Médica pela Universidade Federal de Pernambuco (2003). Especialista em Medicina Interna e Geriatria com área de atuação em Cuidados Paliativos pela Associação Médica Brasileira. Mestrado em Cuidados Paliativos. Doutorado em Saúde Integral, ambos pelo IMIP - Instituto de Medicina Integral Professor Fernando Figueira.

ORCID: https://orcid.org/0000-0003-1130-1098

COORIENTADORA

Rafaella Christine Tenório de Arruda Macêdo

Mestre em Enfermagem pelo Programa de Pós-graduação em Enfermagem da Universidade Federal de Pernambuco (UFPE). Enfermeira do serviço de Oncologia Pediátrica do Instituto de Medicina Integral Professor Fernando Figueira (IMIP). Docente de Enfermagem da Faculdade Pernambucana de Saúde (FPS) e Coordenadora da pós-graduação de Enfermagem em Oncologia da FPS.

ORCID: https://orcid.org/0000-0002-4879-38

ESTUDANTES

Rayne Lyanna Oliveira Dantas

Discente do 6º período do curso de Medicina da Faculdade Pernambucana de Saúde - FPS

https://orcid.org/0009-0004-3122-3204

Beatriz Bronzeado Clericuzi

Discente do 6º período do curso de Medicina da FPS
https://orcid.org/0009-0008-3786-1329.

João Victor Melo Soares

Discente do 10º período do curso de Medicina da FPS
ORCID: https://orcid.org/0009-0001-8741-100X

Lia Sancho Monteiro

Discente do 10º período do curso de Medicina da FPS
https://orcid.org/0000-0001-6765-3960.

Sabrina Vitória das Mercês Cavalcanti
Discente do 10º Período de Medicina na FPS
ORCID: https://orcid.org/0009-0006-9689-6041.

Local de realização da pesquisa: A pesquisa foi realizada no IMIP (Instituto Fernando Figueira).

Os autores negam quaisquer conflitos de interesse no desenvolvimento desta pesquisa.

RESUMO EXPANDIDO

Introdução: A comunicação clínica é componente essencial do cuidado centrado na pessoa e pode ser avaliada por instrumentos estruturados. Com o desenvolvimento dos grandes modelos de linguagem, a inteligência artificial passou a ser investigada para essa finalidade. No entanto, os estudos apresentam instrumentos, contextos e métricas heterogêneos, o que limita a interpretação de sua confiabilidade. **Objetivo:** Analisar o uso da inteligência artificial (IA) na avaliação de habilidades de comunicação humana em saúde e o grau de concordância entre suas avaliações e as de avaliadores humanos. **Métodos:** Revisão narrativa estruturada nas bases PubMed/MEDLINE e BVS/LILACS, com restrição ao idioma inglês. A seleção foi realizada por dois revisores independentes e os resultados foram submetidos a síntese narrativa. **Resultados:** Foram incluídos 25 estudos: em 11, a IA avaliou diretamente a comunicação humana; em 8, a relação com a pergunta de pesquisa foi parcial; e em 6, a IA não avaliou a comunicação. Predominaram análises baseadas em texto ou transcrição. Nos estudos com comparação formal com avaliadores humanos, a concordância variou de elevada (κ = 0,832; ICC = 0,86; F1 = 85,0–92,0%) a baixa (ICC = 0,00–0,35). As métricas empregadas avaliam propriedades distintas e não são diretamente comparáveis. **Conclusão:** A IA demonstrou potencial para avaliar componentes estruturados da comunicação clínica, sobretudo os baseados no conteúdo verbal; contudo, as evidências são limitadas e heterogêneas. No estágio atual, seu uso é mais apropriado como ferramenta complementar, sob supervisão humana, na avaliação formativa e no feedback, não havendo respaldo para a substituição de avaliadores humanos ou para avaliações de alto impacto.

Palavras-chave: Inteligência artificial; Comunicação em saúde; Educação em saúde.

RESUMO

Objetivo: Analisar o uso da inteligência artificial (IA) na avaliação de habilidades de comunicação humana em saúde e o grau de concordância entre suas avaliações e as de avaliadores humanos. **Métodos:** Revisão narrativa estruturada nas bases PubMed/MEDLINE e BVS/LILACS, com restrição ao idioma inglês. A seleção foi realizada por dois revisores independentes e os resultados foram submetidos a síntese narrativa. **Resultados:** Foram incluídos 25 estudos: em 11, a IA avaliou diretamente a comunicação humana; em 8, a relação com a pergunta de pesquisa foi parcial; e em 6, a IA não avaliou a comunicação. Predominaram análises baseadas em texto ou transcrição. Nos estudos com comparação formal com avaliadores humanos, a concordância variou de elevada ($\kappa = 0,832$; ICC = 0,86; F1 = 85,0–92,0%) a baixa (ICC = 0,00–0,35). As métricas empregadas avaliam propriedades distintas e não são diretamente comparáveis. **Conclusão:** A IA demonstrou potencial para avaliar componentes estruturados da comunicação clínica, sobretudo os baseados no conteúdo verbal; contudo, as evidências são limitadas e heterogêneas. No estágio atual, seu uso é mais apropriado como ferramenta complementar, sob supervisão humana, na avaliação formativa e no feedback, não havendo respaldo para a substituição de avaliadores humanos ou para avaliações de alto impacto.

Palavras-chave: Inteligência artificial; Comunicação em saúde; Educação em saúde.

ABSTRACT

Objective: To analyze the use of artificial intelligence (AI) in the assessment of human communication skills in healthcare and the agreement between AI-generated and human assessments. **Methods:** A structured narrative review was conducted in PubMed/MEDLINE and BVS/LILACS, restricted to English-language publications. Two reviewers selected studies independently, and results were synthesized narratively. **Results:** Twenty-five studies were included: in 11, AI directly assessed human communication; in 8, the relationship to the research question was partial; and in 6, AI did not assess communication. Analyses based on text or transcripts predominated. Among studies with formal comparison to human raters, agreement ranged from high (κ=0.832; ICC=0.86; F1=85.0–92.0%) to low (ICC=0.00–0.35). The metrics employed assess distinct properties and are not directly comparable. **Conclusion:** AI showed potential for assessing structured components of clinical communication, particularly those based on verbal content; however, the evidence is limited and heterogeneous. At present, its use is more appropriate as a complementary tool, under human supervision, for formative assessment and feedback, and the evidence does not support replacing human raters or using AI in high-stakes assessment.

Keywords: Artificial intelligence; Health communication; Health education.

I INTRODUÇÃO

A comunicação clínica constitui competência fundamental para o cuidado centrado na pessoa e abrange a obtenção e a transmissão de informações, a construção da relação terapêutica, a expressão de empatia, a organização da consulta e a participação do paciente no processo decisório. Em razão de sua natureza multidimensional e dependente do contexto, a avaliação dessa competência representa um desafio tanto para a educação em saúde quanto para a pesquisa.

Diversos instrumentos foram desenvolvidos para sistematizar essa avaliação, e nem todos mensuram o mesmo construto. O *Consultation and Relational Empathy* (CARE) Measure baseia-se na percepção do paciente acerca da empatia relacional na consulta, com propriedades iniciais de validade e confiabilidade demonstradas na atenção primária.[1] O *Calgary-Cambridge Referenced Observation Guides*, por sua vez, é um referencial estruturado das habilidades de comunicação médico-paciente, que orienta o ensino, a observação e o feedback dessas competências.[2] Assim, enquanto o primeiro reflete a perspectiva do paciente, o segundo se fundamenta na observação por terceiros.

A avaliação estruturada da comunicação depende, em geral, de observadores humanos capacitados, o que demanda treinamento, calibração e tempo, além de estar sujeita à variabilidade entre avaliadores. Nesse contexto, o desenvolvimento do processamento de linguagem natural, do aprendizado de máquina e, mais recentemente, dos modelos de linguagem de grande porte (*large language models* – LLMs) despertou interesse pela automatização parcial desse processo. Na psicoterapia, por exemplo, foram desenvolvidos sistemas de aprendizado de máquina para analisar sessões e fornecer feedback automatizado sobre componentes da entrevista motivacional.[3]

Com a expansão dos LLMs, passaram a ser investigadas estratégias nas quais esses modelos analisam transcrições de interações, aplicam critérios estruturados de avaliação e geram feedback sobre o desempenho dos participantes. Holderried et al. avaliaram, em estudo prospectivo com estudantes de Medicina, um paciente simulado baseado no GPT-4 capaz de fornecer feedback estruturado sobre a anamnese.[4] Ju et al. desenvolveram um avaliador baseado em GPT para aplicar o *Gap-Kalamazoo Communication Skills Assessment Form* (GKCSAF) a transcrições de interações entre terapeutas e pacientes, comparando seus resultados aos de avaliadores humanos.[5] Wang et al. compararam a responsividade do

GKCSAF quando pontuado por especialistas e pelo ChatGPT, verificando a capacidade de ambos de detectar mudanças nas habilidades de comunicação ao longo do tempo.[6] Herschbach et al. confrontaram avaliações geradas por IA e por avaliador humano com o instrumento OPTION-12, em simulação voltada à tomada de decisão compartilhada.[7] Em interações clínicas reais, Lopez-Garcia et al. empregaram LLMs para pontuar a qualidade da comunicação de risco em consultas sobre câncer de próstata.[8]

Contudo, a interpretação dessas evidências exige cautela. Os estudos diferem quanto às populações, aos contextos (clínicos ou simulados), às modalidades de dados fornecidas à IA, aos modelos, aos instrumentos de referência e às estratégias de comparação com avaliadores humanos. Ademais, medidas como coeficiente de correlação intraclasse (ICC), kappa, correlação, acurácia, F1 e responsividade avaliam propriedades distintas e não constituem evidências equivalentes de concordância ou de validade. A natureza da informação disponibilizada ao sistema também é relevante: modelos que recebem apenas transcrições podem avaliar componentes verbais e estruturais da interação, mas não dispõem de elementos prosódicos, gestuais ou faciais. Do mesmo modo, instrumentos concebidos para perspectivas distintas, como a percepção do paciente e a observação por terceiros, não são intercambiáveis quando aplicados por sistemas de IA.

Diante da rápida expansão dessas aplicações e da heterogeneidade metodológica, torna-se necessário caracterizar criticamente como a IA tem sido empregada na avaliação de habilidades de comunicação em saúde, quais instrumentos, dimensões e tipos de dados foram utilizados e que evidências existem de comparação com avaliadores humanos. Essa síntese pode contribuir para delimitar as possibilidades e as limitações atuais da avaliação automatizada da comunicação clínica.

II OBJETIVOS

Objetivo geral

Analisar o uso da inteligência artificial na avaliação de habilidades de comunicação humana em contextos clínicos e de educação em saúde, bem como o grau de concordância entre suas avaliações e as de avaliadores humanos.

Objetivos específicos

Identificar os modelos e as técnicas de inteligência artificial empregados na avaliação da comunicação humana;

Caracterizar os contextos, as populações e as modalidades de dados analisados (texto, áudio, vídeo ou dados multimodais);

Identificar os instrumentos, escalas, listas de verificação ou rubricas que orientaram a avaliação automatizada e as dimensões da comunicação avaliadas;

Analisar as estratégias de comparação entre as avaliações produzidas pela IA e por avaliadores humanos, considerando as métricas de concordância, confiabilidade, desempenho ou responsividade reportadas; e

Identificar limitações metodológicas e lacunas de evidência relevantes para aplicações futuras da IA na avaliação da comunicação clínica.

III MÉTODOS

Delineamento: Trata-se de revisão narrativa estruturada da literatura sobre o uso da inteligência artificial na avaliação de habilidades de comunicação humana em contextos clínicos e de educação em saúde. A revisão foi conduzida com estratégia de busca previamente definida, triagem independente dos registros e extração estruturada das características metodológicas e dos resultados dos estudos selecionados.

Bases de dados e filtros: Foram realizadas buscas nas bases PubMed/MEDLINE e Biblioteca Virtual em Saúde (BVS/LILACS), com aplicação de filtro de idioma (inglês). *[Confirmar e acrescentar os demais filtros utilizados, como período, tipo de publicação e data da busca.]* Como todos os descritores foram empregados em inglês, estudos publicados em português ou espanhol podem não ter sido recuperados, o que constitui limitação da estratégia de busca.

Estratégia de busca: Foram combinados três blocos de descritores, unidos internamente pelo operador booleano **OR** e entre si pelo operador **AND**:

Bloco 1 (inteligência artificial): "artificial intelligence" OR "machine learning" OR "large language model" OR LLM OR ChatGPT OR GPT OR "generative AI"

Bloco 2 (comunicação clínica): "communication skills" OR "clinical communication" OR "physician-patient communication" OR "patient-physician communication" OR empathy

Bloco 3 (contexto ou processo avaliativo): simulation OR “simulated patient” OR “standardized patient” OR OSCE OR assessment OR evaluation OR scoring OR feedback

Assim, a estrutura geral da busca foi: (Bloco 1) **AND** (Bloco 2) **AND** (Bloco 3). As estratégias foram adaptadas às características de indexação de cada base, e o número de registros recuperados e a estratégia utilizada em cada fonte foram preservados, a fim de garantir a rastreabilidade do processo.

Critérios de elegibilidade: Foram considerados elegíveis os estudos que investigaram o uso de inteligência artificial, aprendizado de máquina, processamento de linguagem natural ou modelos de linguagem de grande porte para avaliar, analisar, classificar ou pontuar aspectos da comunicação de estudantes, residentes ou profissionais de saúde, ou para gerar feedback com base nesse desempenho. Foram admitidos contextos clínicos reais ou simulados, incluindo interações com pacientes simulados ou padronizados, e diferentes modalidades de entrada (texto, áudio, vídeo e dados multimodais). Estudos nos quais a IA atuou exclusivamente como interlocutor, paciente virtual, ferramenta de preparação da consulta ou geradora de conteúdo, sem analisar a comunicação do participante humano, não foram considerados evidência principal. Protocolos, artigos conceituais, perspectivas e publicações sem dados empíricos foram separados da síntese analítica principal.

Seleção dos estudos: Os registros identificados nas diferentes bases foram reunidos em uma base única e as duplicatas foram removidas. Em seguida, dois revisores realizaram, de forma independente, a triagem por título e resumo, classificando cada registro como “incluir”, “excluir” ou “dúvida”. As discordâncias e os registros em dúvida foram submetidos à adjudicação. Os estudos potencialmente elegíveis foram encaminhados à leitura do texto completo, etapa na qual a elegibilidade foi reavaliada com base na pergunta da revisão e na função efetivamente desempenhada pela IA. O fluxo de seleção está representado na Figura 1.

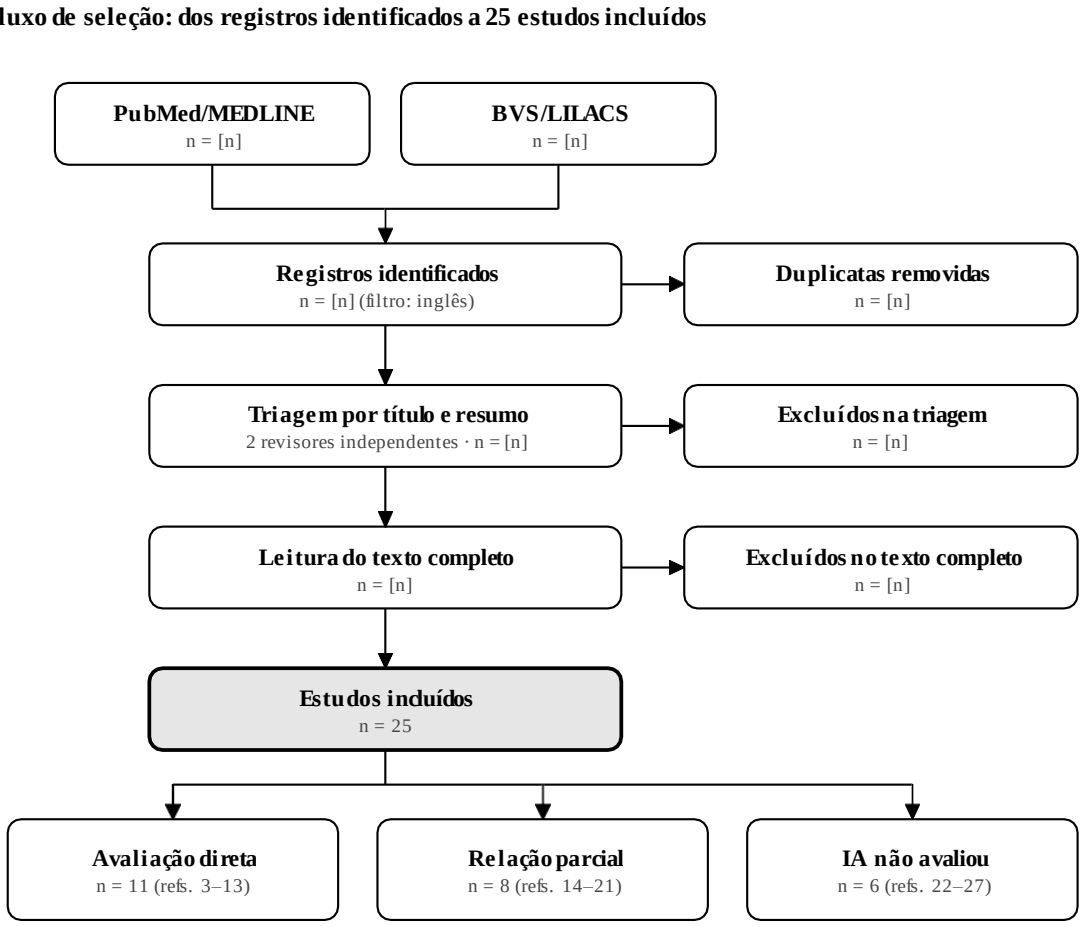
Para fins de síntese, os estudos foram organizados em dois estratos analíticos. O núcleo de validação (A1) compreendeu os estudos em que a IA avaliou a comunicação humana e foi apresentada comparação ou evidência de validação em relação a avaliadores humanos ou a outro padrão de referência pertinente. O núcleo de aplicação (A2) reuniu os estudos em que a IA analisou diretamente componentes da comunicação humana, sem comparação IA-humano suficientemente robusta ou como parte de uma aplicação mais ampla. Estudos sobre

treinamento de comunicação, pacientes virtuais, avaliação da comunicação produzida pela própria IA ou temas conceitualmente adjacentes foram mantidos como evidência contextual.

Extração dos dados: Utilizou-se formulário padronizado para a extração das seguintes informações: autoria e ano de publicação; país; objetivo; delineamento; contexto e população; número de participantes e de interações analisadas; modalidade de dado fornecida à IA; modelo e versão; função desempenhada pela IA; instrumento utilizado; dimensão comunicacional avaliada; geração de escores ou feedback; comparação com avaliadores humanos (número, treinamento e cegamento); método estatístico; resultados e intervalos de confiança; reprodutibilidade; e limitações. Uma ferramenta de IA generativa foi empregada como apoio à pré-extração, sem ser considerada fonte definitiva: as informações foram conferidas por revisores humanos no texto completo e as divergências foram corrigidas com base na fonte primária. Informações não apresentadas pelos autores foram registradas como "não informado", sem imputação ou inferência. Após a leitura do texto completo, os revisores classificaram cada estudo como "sim", "parcialmente" ou "não" quanto ao uso da IA para avaliar a comunicação de uma pessoa, acompanhado de justificativa.

Síntese e análise dos dados: Em razão da heterogeneidade esperada entre populações, contextos, tecnologias, modalidades de entrada, instrumentos e medidas de desempenho, não foi planejada combinação quantitativa dos resultados. A síntese foi descritiva e narrativa, e as métricas foram interpretadas conforme a propriedade avaliada: medidas de concordância ou confiabilidade, de associação, de desempenho classificatório, de erro e índices de responsividade foram apresentadas separadamente, evitando-se sua interpretação como equivalentes.

Figura 1. Fluxograma de identificação e seleção dos estudos



Nota: os valores indicados como [n] devem ser substituídos pelos números do fluxograma apresentado no slide. Os 25 estudos incluídos foram classificados em 11, 8 e 6, conforme o papel desempenhado pela IA.

IV RESULTADOS E DISCUSSÃO

Caracterização dos estudos

Foram analisados 25 estudos. Em 11, a IA avaliou diretamente a comunicação de profissionais ou estudantes de saúde (refs. 3–13). Em 8, a relação com a pergunta da revisão foi parcial, pois a IA classificou textos sobre comunicação, aplicou modelos previamente validados ou atuou como ferramenta auxiliar de análise (refs. 14–21). Nos 6 restantes, a IA não avaliou a comunicação de nenhuma pessoa (refs. 22–27).

Entre os estudos de avaliação direta, predominaram entradas em texto ou transcrição (refs. 4–8, 11, 13). Um estudo utilizou áudio processado por reconhecimento automático de fala (ref. 3) e três empregaram vídeo ou dados multimodais (refs. 9, 10, 12). Os instrumentos incluíram escalas validadas (OPTION-12, GKCSAF, LUCAS e MISC/MITI, este último

empregado no treinamento do modelo), listas de verificação ou rubricas elaboradas pelos próprios autores (refs. 4, 8, 13) e métricas comportamentais não clínicas (refs. 9, 10, 12).

Heterogeneidade das métricas

As métricas empregadas foram heterogêneas, e cada uma responde a uma pergunta distinta. Por essa razão, não devem ser agregadas nem resumidas genericamente como "boa concordância". O Quadro 1 apresenta o que cada métrica avalia e sua interpretação usual.

Quadro 1. Métricas utilizadas nos estudos e seu significado

Métrica	O que avalia	Estudos	Interpretação
Kappa (κ)	Concordância entre duas avaliações categóricas, descontado o acaso	4, 15	Acima de 0,80: quase perfeita; 0,21–0,40: razoável[28]
ICC	Confiabilidade de escores numéricos entre avaliadores	3, 5, 7	Abaixo de 0,50: confiabilidade pobre[29]
F1, acurácia e AUC	Desempenho classificatório em relação a uma referência	8, 13, 16	Expressam acertos, e não concordância entre avaliadores
MAE% (erro absoluto médio)	Magnitude média da diferença entre os escores	5	Pode ser aceitável mesmo com ICC baixo
Responsividade (d, SRM)	Capacidade de detectar mudança ao longo do tempo	6	Não demonstra concordância no mesmo encontro[32]

O estudo de Ju et al. ilustra essa distinção: o erro médio foi aceitável, ao passo que o ICC permaneceu próximo de zero.[5] Tal achado é coerente com a sensibilidade do kappa e do ICC à baixa variabilidade dos escores.[29,31] De modo semelhante, no estudo de Wang et al., o ChatGPT e os especialistas apresentaram responsividade comparável, o que não implica atribuição de escores equivalentes ao mesmo encontro.[6] Ademais, um ICC elevado pode coexistir com viés sistemático: Herschbach et al. observaram que a IA atribuiu notas mais altas que o avaliador humano,[7] e Alhaskir et al. verificaram médias próximas, porém com variabilidade distinta.[11] Por essa razão, a concordância absoluta deve ser examinada por métodos como o de Bland-Altman, e não inferida a partir da correlação ou da proximidade entre médias.[30]

Estudos com comparação com avaliadores humanos

O Quadro 2 sintetiza os principais estudos que compararam a IA a avaliadores humanos, evidenciando a heterogeneidade entre eles quanto a dados, instrumentos, métricas e limitações.

Quadro 2. Principais estudos com comparação entre IA e avaliadores humanos

Estudo	Dado e amostra	Instrumento	Métrica e resultado	Ponto de atenção
Holderried et al.[4]	Transcrição; 106 estudantes (1.894 pares pergunta-resposta)	Lista de verificação própria (45 categorias)	$\kappa = 0,832$ (IC95% 0,816–0,848)	Um avaliador humano; 8 categorias com $\kappa < 0,64$
Imel et al.[3]	Áudio; 21 terapeutas	MISC/MITI	ICC de 0,80 (perguntas fechadas) a 0,23 (empatia); média 0,48	Um avaliador; concordância baixa para empatia
Lopez-Garcia et al.[8]	Transcrição; 487 frases de 20 consultas reais	Rubrica própria (0–5)	F1 = 85,0–92,0; entre humanos, 83,0–85,4	Dependente da estratégia de prompt
Herschbach et al.[7]	Transcrição; 58 participantes	OPTION-12	ICC = 0,86 (0,633 antes do ajuste do prompt)	Um especialista; IA atribuiu notas mais altas
Ju et al.[5]	Transcrição; 80 interações	GKCSAF	ICC = 0,00–0,35; kappa ponderado ≤ 0,38; MAE% 12,2–21,0	Concordância pobre em todos os domínios
Wang et al.[6]	Texto; 160 interações de 80 estudantes	GKCSAF	Responsividade: d = 0,47 (IA) e 0,45 (especialistas)	Concordância ponto a ponto não avaliada
Maicher et al.[13]	Diálogos; 20	Lista de verificação própria	Acurácia de 87% (IA) e 90% (3 humanos)	Amostra reduzida
Alhaskir et al.[11]	Texto; 4 simulações	LUCAS	Média de 13,9 (IA) e 12,8 (humanos)	Sem kappa ou ICC; quatro simulações

Desempenho segundo a dimensão avaliada

O desempenho da IA foi mais consistente em dimensões estruturadas e baseadas em conteúdo do que em dimensões relacionais da comunicação. A concordância com avaliadores humanos foi mais elevada na completude da anamnese ($\kappa = 0,832$),[4] na coleta de informações,[13] na comunicação de risco pontuada por rubrica (F1 de 85,0 a 92,0)[8] e em comportamentos quantificáveis, como perguntas fechadas (ICC = 0,80).[3] Nas dimensões relacionais, os resultados foram inferiores: a empatia apresentou ICC de 0,23,[3] os nove domínios do GKCSAF exibiram kappa ponderado $\leq 0,38$[5] e a empatia pontuada por aprendizado de máquina não se associou ao desfecho clínico.[18]

Padrão semelhante foi observado quando a IA atuou apenas como apoio ao profissional, com ganhos na organização das informações, mas sem benefício, ou com prejuízo, em aspectos humanísticos e em itens específicos de comunicação.[23,26] Em um dos estudos, o cuidado humanístico reduziu-se nos modelos totalmente ajustados (B = −3,414; IC95% −5,723 a −1,105)[23] e, no apoio à triagem telefônica, observou-se discreta melhora da qualidade médica, sem alteração da qualidade geral da comunicação e com piora em dois itens específicos.[26]

Limitações metodológicas dos estudos

Os resultados mais próximos aos dos avaliadores humanos provieram, em geral, de amostras pequenas e de comparadores frágeis. Diversos estudos compararam a IA a um único avaliador humano (refs. 3, 4, 7), analisaram apenas 20 diálogos[13] ou quatro simulações.[11] Nessas condições, a estimativa de concordância depende da consistência de um único avaliador, cujo padrão de referência pode apresentar confiabilidade apenas moderada, como no estudo em que os avaliadores humanos concordaram entre si com Fleiss $\kappa = 0{,}44$.[16] Os resultados também dependeram da calibração do prompt: em um estudo, o ICC passou de 0,633 para 0,86 após refinamento manual,[7] e em outro o F1 variou conforme a estratégia de prompting.[8]

A modalidade de entrada também limitou a validade dos achados. As validações comparativas com humanos concentraram-se em texto e transcrições (refs. 4, 5, 7, 8, 13). Os estudos com vídeo ou dados multimodais mensuraram comportamentos não verbais ou emoções sem validar essas medidas frente a observadores humanos. Kobayashi et al. observaram aumento do contato visual de 5,7% para 27,4% do tempo de consulta após treinamento, sem comparação com observadores;[9] Lambrakos et al. mencionaram verificação manual não quantificada;[10] e Huang et al. não relataram a acurácia do sistema de reconhecimento de emoções.[12] O único estudo com áudio e comparação formal apresentou concordância modesta.[3] Assim, os componentes não verbais e afetivos permanecem os menos validados.

Estudos sem relação direta com a pergunta da revisão

Entre os oito estudos com relação parcial, quatro utilizaram a IA para analisar textos ou dados sobre comunicação, e não o desempenho comunicativo propriamente dito. Lou et al. associaram fatores do médico à satisfação de pacientes padronizados por meio de dois modelos, que produziram associações contraditórias para as mesmas variáveis.[14] Ahmad et al. empregaram o ChatGPT para classificar comentários de feedback segundo as competências do ACGME, com concordância de 90% ($\kappa = 0,94$) em 60 entradas, sendo a comunicação interpessoal apenas uma das seis competências.[15] Booth et al. classificaram 25.714 frases de avaliações narrativas por processamento de linguagem natural, com área sob a curva entre 0,59 e 0,88 e concordância apenas moderada entre os avaliadores humanos que definiram a referência.[16] Dois estudos do mesmo grupo aplicaram, em larga escala, modelos previamente validados, sem nova comparação com humanos.[17,18] Nos demais, a IA exerceu papel instrumental: apoio à análise temática de relatos de pacientes,[19] robô-paciente cujo ICC de 0,87 refere-se à concordância entre dois avaliadores humanos[20] e pré-triagem de entrevistas qualitativas com validação humana.[21]

Nos seis estudos restantes, a IA não avaliou a comunicação: gerou conteúdo educativo sobre comunicação de más notícias,[22] apoiou a preparação ou a decisão do profissional[23,25,26] ou mensurou apenas percepções sobre a tecnologia.[24,27] Esses trabalhos informam sobre aceitabilidade e efeitos educacionais, mas não sobre a acurácia da IA como avaliadora, e sua separação evita a superestimação da evidência disponível. Do mesmo modo, a aplicação de modelos em larga escala[17,18] não substitui a validação da concordância em cada novo contexto.

Síntese

Em conjunto, a IA apresentou concordância de moderada a alta com avaliadores humanos em tarefas textuais e estruturadas e concordância baixa ou não demonstrada nas dimensões relacionais e não verbais, em um corpo de evidência limitado por amostras pequenas, comparadores humanos únicos, métricas heterogêneas e dependência do prompt.

V CONCLUSÃO

A inteligência artificial mostrou potencial como avaliadora de habilidades de comunicação clínica, mas as evidências disponíveis sustentam esse uso apenas em parte e em condições específicas. Nas tarefas estruturadas e baseadas no conteúdo verbal da interação, como a completude da anamnese, a coleta de informações, a comunicação de risco pontuada por rubrica e a identificação de comportamentos quantificáveis, os sistemas apresentaram desempenho próximo ao de avaliadores humanos, expresso por medidas distintas (kappa, ICC, F1 e acurácia) que não são equivalentes entre si. Nas dimensões relacionais, como a empatia, a concordância foi baixa. Nos componentes não verbais e afetivos, analisados por áudio e vídeo, a concordância com observadores humanos não foi demonstrada.

A robustez desses achados é limitada. A maior parte dos estudos teve amostras pequenas, comparou a IA a um único avaliador humano e utilizou instrumentos próprios ou adaptados. Nenhum estudo do núcleo principal empregou o CARE ou o guia Calgary-Cambridge, e nenhum avaliou a reprodutibilidade das avaliações geradas pela IA em repetições. O desempenho também dependeu do modelo e da estratégia de prompt. Além disso, 14 dos 25 estudos não avaliaram diretamente a comunicação humana, empregando a IA como simuladora, geradora de conteúdo, apoio à decisão ou ferramenta de análise, distinção necessária para evitar a superestimação da evidência.

No estágio atual, a IA pode ser considerada ferramenta complementar, sob supervisão humana, para a avaliação formativa e o feedback sobre componentes estruturados da comunicação, em contextos educacionais e clínicos. Seu uso em avaliações somativas ou de alto impacto, ou como substituta do avaliador humano, não é sustentado pelas evidências disponíveis.

Estudos futuros devem utilizar instrumentos validados e adequados à perspectiva avaliada, considerando que medidas baseadas na percepção do paciente, como o CARE, e guias de observação por terceiros, como o Calgary-Cambridge, não são intercambiáveis. Devem, ainda, incluir mais de um avaliador humano treinado e cegado, relatar medidas de concordância absoluta com intervalos de confiança e examinar a reprodutibilidade dos modelos. Por fim, é necessário abordar as dimensões relacionais e não verbais da comunicação e testar essas ferramentas em diferentes populações, contextos e idiomas,

incluindo estudos conduzidos no Brasil e em língua portuguesa, que não foram identificados nesta revisão, possibilidade em parte relacionada à restrição ao idioma inglês.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Mercer SW, Maxwell M, Heaney D, Watt GCM. The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. Fam Pract. 2004;21(6):699-705. doi:10.1093/fampra/cmh621.
2. Kurtz SM, Silverman JD. The Calgary-Cambridge Referenced Observation Guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. Med Educ. 1996;30(2):83-89. doi:10.1111/j.1365-2923.1996.tb00724.x.
3. Imel ZE, Pace BT, Soma CS, Tanana M, Hirsch T, Gibson J, et al. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. Psychotherapy (Chic). 2019;56(2):318-328. doi:10.1037/pst0000221.
4. Holderried F, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ. 2024;10:e59213. doi:10.2196/59213.
5. Ju YJ, Wang YC, Lee SC, Liu CH, Lee ML, Hou CY, et al. Development and validation of a GPT-based rater for assessing communication skills using the Gap-Kalamazoo Communication Skills Assessment Form. Med Teach. 2026;48(1):93-104. doi:10.1080/0142159X.2025.2532783.
6. Wang YC, et al. Detecting change in clinical communication skills: responsiveness of the Gap-Kalamazoo communication skills assessment form scored by experts and ChatGPT. Patient Educ Couns. 2026. doi:10.1016/j.pec.2026.109829.
7. Herschbach L, Festl-Wietek T, Sonanini A, Holderried F, Albrecht T, Herrmann B, et al. AI-simulated patients for training shared decision-making: feasibility study in medical education. JMIR Med Educ. 2026;12:e100467. doi:10.2196/100467.
8. Lopez-Garcia G, Xu D, Luu M, Zheng R, Daskivich TJ, Gonzalez-Hernandez G. Scoring physician risk communication in prostate cancer using large language models. Pac Symp Biocomput. 2026;31:71-84. doi:10.1142/9789819824755_0006.
9. Kobayashi M, Katayama M, Hayashi T, Hashiyama T, Iyanagi T, Une S, et al. Effect of multimodal comprehensive communication skills training with video analysis by artificial intelligence for physicians on acute geriatric care: a mixed-methods study. *BMJ Open*. 2023;13(3):e065477. doi:10.1136/bmjopen-2022-065477.
10. Lambrakos LK, et al. Enhancing patient acceptance of ICD implantation through structured shared decision making: conversation is key. J Interv Card Electrophysiol.

2024. doi:10.1007/s10840-024-01850-7.

11. Alhaskir M, Haven H, Langer M, Peters T, Hertwig M, Unterkofler J, et al. Feasibility of LLM-assisted communication feedback in simulation-based medical education: a pilot study. Stud Health Technol Inform. 2026;336:2140-2144. doi:10.3233/SHTI260639.
12. Huang CW, et al. Emotion recognition in doctor-patient interactions from real-world clinical video database: initial development of artificial empathy. Comput Methods Programs Biomed. 2023;233:107480. doi:10.1016/j.cmpb.2023.107480.
13. Maicher KR, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach. 2019;41(9):1053-1059. doi:10.1080/0142159X.2019.1616683.
14. Lou Z, et al. Examining how physician factors influence patient satisfaction during clinical consultations about cancer prognosis and pain. PEC Innov. 2022;1:100017. doi:10.1016/j.pecinn.2022.100017.
15. Ahmad SA, et al. Evaluating resident feedback using a large language model: are we missing core competencies? Laryngoscope. 2025;135(11):4119-4124. doi:10.1002/lary.32368.
16. Booth GJ, Ross B, Cronin WA, McElrath A, Cyr KL, Hodgson JA, et al. Competency-based assessments: leveraging artificial intelligence to predict subcompetency content. Acad Med. 2023;98(4):497-504. doi:10.1097/ACM.0000000000005115.
17. Zhang X, et al. You never know what you are going to get: large-scale assessment of therapists' supportive counseling skill use. Psychotherapy (Chic). 2023;60(2):149-158. doi:10.1037/pst0000460.
18. Zhang X, Goldberg SB, Baldwin SA, Tanana MJ, Weitzman LM, Narayanan SS, et al. Association of machine-learning-rated supportive counseling skills with psychotherapy outcome. J Consult Clin Psychol. 2025;93(2):110-119. doi:10.1037/ccp0000935.
19. Stage MA, et al. “Having providers who are trained and have empathy is life-saving”: improving primary care communication through thematic analysis with ChatGPT and human expertise. PEC Innov. 2025;6:100371. doi:10.1016/j.pecinn.2024.100371.
20. Wang X, et al. Study on the training effect of SIMROID robot in improving medical students' patient-clinician communication and behavior management skills. J Dent

Educ. 2026. doi:10.1002/jdd.70076.

21. Li S, et al. Exploring dynamic changes in empathic capability among postgraduate nursing students during clinical internships: an artificial intelligence-assisted qualitative study with human-in-the-loop validation. Nurse Educ. 2026. doi:10.1097/NNE.0000000000002188.
22. Armstrong ME, et al. Can artificial intelligence help with the development of generic clinical skills when breaking bad news? A quantitative evaluation. Clin Med (Lond). 2026;26(3):100562. doi:10.1016/j.clinme.2026.100562.
23. Wei Y, et al. The double-edged sword of generative AI in dermatology: a multicomponent cross-sectional study on physician burnout, patient satisfaction, and communication quality. Front Med (Lausanne). 2026;13:1875075. doi:10.3389/fmed.2026.1875075.
24. Saad M, et al. Exploring the use of artificial intelligence-enhanced virtual reality in teaching and evaluating communication skills in urology: a pilot study. Cureus. 2026. doi:10.7759/cureus.109258.
25. Gao C, et al. The VALUE of AI-guided communication: enhancing shared decision-making in metabolic bariatric surgery consultations through a metacognitive framework. Obes Surg. 2026. doi:10.1007/s11695-026-08760-9.
26. Fotland SLS, Berge A, Zakariassen E, Midtbø V, Baste V, Fonnes G, et al. Impact of a prototype combining recommender functionality with structured documentation on operator performance in calls to medical communication centers: quasi-experimental feasibility study. JMIR Form Res. 2026;10:e87082. doi:10.2196/87082.
27. Macdonald M, et al. Effective communication training to support person- and whānau-centred care: what is needed and why isn't it happening? N Z Med J. 2026. doi:10.26635/6965.7045.
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174.
29. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012.
30. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-310.

31. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-549. doi:10.1016/0895-4356(90)90158-L.
32. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-745. doi:10.1016/j.jclinepi.2010.02.006.