UNIVERSIDADE CHRISTUS

MESTRADO PROFISSIONAL EM ENSINO NA SAÚDE

MARCELO LIMA GONZAGA

AVALIAÇÃO DO FEEDBACK POR FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA NO ENSINO MÉDICO

FORTALEZA

2026

MARCELO LIMA GONZAGA

AVALIAÇÃO DO FEEDBACK POR FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA NO ENSINO MÉDICO

Dissertação apresentada à Universidade Christus para obtenção de qualificação de Mestrado em Ensino na Saúde e Tecnologias Educacionais. Área de concentração: Ensino na Saúde. Linha de pesquisa: Avaliação do Ensino e Aprendizagem em Saúde.

Orientador: Prof. Dr. Hermano Alexandre Lima Rocha

Coorientador: Prof Dr. Marcos Kubrusly

FORTALEZA

2026

FICHA CATALOGRÁFICA

Dados Internacionais de Catalogação na Publicação
Universidade Christus - Unichristus
Gerada automaticamente pelo Sistema de Elaboração de Ficha Catalográfica da
Universidade Christus - Unichristus, com dados fornecidos pelo(a) autor(a)

L732a Lima Gonzaga, Marcelo.
Avaliação do feedback por ferramentas de inteligência artificial generativa no ensino médico / Marcelo Lima Gonzaga. - 2026.
64 f. : il. color.

Dissertação (Mestrado) - Universidade Christus - Unichristus, Mestrado em Ensino na Saúde e Tecnologias Educacionais, Fortaleza, 2026.
Orientação: Prof. Dr. Hermano Alexandre Lima Rocha.
Coorientação: Prof. Dr. Marcos Kubrusly.
Área de concentração: Ensino em Saúde.

1. inteligência artificial generativa. 2. ensino médico. 3. avaliação no ensino superior. I. Título.

CDD 610.7

MARCELO LIMA GONZAGA

AVALIAÇÃO DO FEEDBACK POR FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA NO ENSINO MÉDICO

Dissertação apresentada a Universidade Christus de Fortaleza para obtenção do título de Mestrado em Ensino na Saúde e Tecnologias Educacionais. Área de concentração: Ensino na Saúde. Linha de pesquisa: Avaliação do Ensino e Aprendizagem Em Saúde.

Orientador: Prof. Dr. Hermano Alexandre Lima
Coorientador: Prof. Dr. Marcos Kubrusly

Aprovado em: ____/____/_____

BANCA EXAMINADORA

_____

Prof. Dr. Hermano Alexandre Lima Rocha

Orientador

(Universidade Christus)

_____

Prof. Dr. Marcos Kubrusly
Coorientador

(Universidade Christus)

_____

Prof. Dr. Diego Bastos

(Universidade Christus/Universidade Federal do Ceará)

_____

Prof. Dra. Juliana Arruda

Universidade Estadual do Ceará (UECE)

AGRADECIMENTOS

Agradeço, primeiramente, a Deus pela condução serena deste percurso e pela oportunidade de crescimento pessoal e profissional ao longo desta jornada. Aos meus orientadores, professores Dr. Marcos Kubrusly e Dr. Hermano Rocha, expresso minha profunda gratidão pela confiança, pelo rigor científico e pela generosidade intelectual com que conduziram este trabalho. Seus ensinamentos ultrapassam as páginas deste estudo e permanecerão como referência em minha formação.

À minha esposa, Natália, pelo apoio incondicional, pela compreensão nos momentos de ausência e por caminhar ao meu lado com leveza e firmeza, tornando este percurso possível e mais significativo.

Aos meus pais e irmãos, pelo incentivo constante, pelos valores transmitidos e pelo alicerce sólido que sempre representaram em minha trajetória.

Aos meus filhos, Eduardo, com seus dois anos de vida, e Maria Luisa, ainda em gestação, por serem fonte de motivação, amor e propósito em tudo o que realizo.

Aos alunos de iniciação científica, pela dedicação, pelo comprometimento e pela contribuição valiosa no desenvolvimento deste estudo, tornando-o mais rico e consistente.

À Universidade Christus, pela formação acadêmica de excelência e pelo ambiente que favorece o desenvolvimento científico, crítico e humano.

A todos que, de alguma forma, contribuíram para a realização deste trabalho, deixo aqui o meu sincero agradecimento.

RESUMO

A IA generativa tem ganhado destaque por sua capacidade de criar conteúdos realistas e de alta qualidade em áreas como geração de imagens, síntese de texto e modelos de aprendizagem. Na educação, esses modelos podem revolucionar o ensino, oferecendo novos recursos para complementar métodos tradicionais e facilitar a compreensão de conceitos complexos. No contexto da educação médica, a IA generativa permite a criação de imagens médicas sintéticas e simulações interativas, proporcionando aos estudantes uma vasta gama de casos realistas para estudo e prática, sem os riscos associados a pacientes reais. Essa abordagem pode melhorar significativamente a experiência de aprendizagem, oferecendo uma prática clínica segura e realista. O ChatGPT, uma ferramenta de IA, tem potencial para fornecer feedback educacional detalhado e personalizado, auxiliando na análise de dados, na síntese de informações e na resolução de problemas. A metodologia inclui um estudo transversal, com abordagem quantitativa, envolvendo estudantes de diversos semestres e a coleta de dados sobre a percepção dos alunos em relação ao feedback recebido. Estudantes de Medicina de uma universidade de Fortaleza foram submetidos a avaliações teóricas de múltipla escolha e receberam feedback tanto de IA quanto de professores, a partir de suas respostas. O projeto visou avaliar a eficácia dos feedbacks gerados por IA em comparação com os fornecidos por professores humanos. A pesquisa analisa a qualidade e a adequação desses feedbacks, sua conformidade com expectativas teóricas e sua receptividade pelos estudantes, por meio de questionários aplicados.

Os resultados demonstraram que os feedbacks produzidos pela inteligência artificial apresentaram maior especificidade e maior nível de detalhamento, especialmente nos estudantes dos primeiros semestres do curso médico, evidenciando potencial para auxiliar no desenvolvimento inicial do aprendizado conceitual e cognitivo. Entretanto, nos estudantes do sétimo semestre, observou-se maior relevância da sensibilidade humana na construção do feedback educacional, sobretudo em aspectos relacionados à interpretação das dificuldades individuais, à complexidade do raciocínio clínico e à dimensão interpessoal do processo formativo. Observou-se elevada receptividade dos estudantes em relação aos feedbacks gerados por inteligência artificial, com avaliações predominantemente positivas acerca da clareza, capacidade explicativa e potencial de contribuição para o aprendizado. O intuito final foi o desenvolvimento de um manual para a utilização de feedbacks gerados por IA, contribuindo com uma abordagem mais eficiente e adaptativa para o ensino e a aprendizagem.

Palavras-chave: inteligência artificial generativa; ensino médico; avaliação no ensino superior.

ABSTRACT

Generative artificial intelligence (AI) has gained increasing prominence due to its ability to produce realistic and high-quality content in areas such as image generation, text synthesis, and advanced learning models. In the field of education, these technologies have the potential to transform traditional teaching methods by providing innovative resources capable of complementing conventional pedagogical approaches and facilitating the understanding of complex concepts. Within the context of medical education, generative AI enables the development of synthetic medical images and interactive simulations, offering students access to a wide range of realistic clinical scenarios for study and practice without the risks associated with real patients. This approach may significantly enhance the learning experience by providing a safe and realistic environment for clinical training. ChatGPT, an artificial intelligence tool based on large language models, demonstrates potential to deliver detailed and personalized educational feedback, supporting data analysis, information synthesis, and problem-solving processes. The present study employed a cross-sectional design with a quantitative approach, involving medical students from different semesters and the collection of data regarding students' perceptions of the feedback received. Medical students from a university in Fortaleza, Brazil, underwent multiple-choice theoretical assessments and subsequently received feedback generated both by artificial intelligence and by human professors, based on their responses. The study aimed to evaluate the effectiveness of AI-generated feedback in comparison with feedback provided by human educators. Furthermore, the research analyzed the quality and pedagogical adequacy of these feedbacks, their alignment with theoretical expectations, and their acceptance among students through structured questionnaires. The findings demonstrated that feedback produced by artificial intelligence exhibited greater specificity and a higher level of detail, particularly among students in the early semesters of medical school, highlighting its potential to support the initial development of conceptual and cognitive learning. However, among seventh-semester students, human sensitivity emerged as a more relevant component in the construction of educational feedback, especially regarding the interpretation of individual learning difficulties, the complexity of clinical reasoning, and the interpersonal dimension of the educational process. Additionally, students demonstrated high receptiveness toward AI-generated feedback, with predominantly positive evaluations concerning clarity, explanatory capacity, and contribution to learning. The final outcome of the project was the development of a practical guide for the use of AI-generated feedback in medical education, contributing to a more efficient and adaptive approach to teaching and learning processes.

Keywords: generative artificial intelligence; medical education; assessment in higher education.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IA	Inteligência Artificial
IAG	Inteligência Artificial Generativa
TCLE	Termo de Consentimento Livre e Esclarecido
TIC	Tecnologias de Informação e Comunicação
LIT	Laboratório de Inovações Tecnológicas
MESTED	Mestrado em Ensino na Saúde e Tecnologias Educacionais
Unichristus	Universidade Christus

SUMÁRIO

1. INTRODUÇÃO

Desde a origem da IA, pesquisadores e especialistas reportam sobre o seu grande impacto futuro na humanidade (GAŠEVIĆ; SIEMENS; SADIQ, 2023). Os modelos de inteligência artificial generativa têm recebido bastante destaque nos últimos anos devido à sua habilidade em produzir resultados realistas e de alta qualidade em diversos campos, incluindo geração de imagens, síntese de texto, modelos de aprendizado e confecção de feedbacks (CHAO, 2022). Ela pode ser definida como um conjunto de modelos computacionais capazes de aprender padrões a partir de grandes volumes de dados e gerar novos conteúdos com coerência estatística e semântica, distinguindo-se das abordagens tradicionais de IA voltadas à classificação ou predição (GOODFELLOW *et al.*, 2014). Dentro do âmbito empresarial e industrial, os modelos de IA generativa têm encontrado aplicação em diversas áreas, como manufatura inteligente e sistemas de inteligência empresarial (LI *et al.*, 2018).

O ensino superior tem sido influenciado pelo surgimento veloz e pelo crescimento exponencial na capacidade e no uso da inteligência artificial generativa (IA) (LODGE; THOMPSON; CORRIN, 2023). Sua utilização pode desempenhar um papel crucial no avanço dos recursos educacionais na área da educação médica entre os estudantes de graduação, com potencial de complementar os clássicos métodos de ensino e aprimorar a compreensão de conceitos médicos tradicionais mais complexos pelos estudantes (SAVAGE, 2021).

O ChatGPT consiste em um modelo de linguagem baseado em inteligência artificial generativa, desenvolvido a partir da arquitetura *Generative Pre-trained Transformer*. Pode gerar respostas ao receber um *prompt* que delineia a tarefa, por exemplo, por meio de uma instrução ou consulta redigida por um usuário (LODGE; THOMPSON; CORRIN, 2023). Faz-se útil em encontrar artigos ou livros relevantes para aprimorar o conhecimento dos estudantes, auxiliando – por exemplo – na análise e interpretação de dados. Além disso, possui o potencial de impulsionar a

aprendizagem atual sobre a geração de feedback educacional na forma de texto para tarefas executadas previamente pelos estudantes (THURLINGS *et al.*, 2012). Os alunos poderiam inserir seus trabalhos escritos no ChatGPT e solicitar feedback sobre aspectos como a fluidez de um argumento ou estrutura gramatical (LODGE; THOMPSON; CORRIN, 2023), com o uso crítico, reflexivo e eticamente responsável das ferramentas de inteligência artificial, capacitando os estudantes para sua utilização consciente, criteriosa e alinhada aos princípios da prática médica e da formação humanística. Essa realidade pode resultar na concepção de soluções educacionais inovadoras que incorporam tecnologias de IA ao currículo médico, proporcionando, em última instância, melhorias no ensino em saúde (PARDO *et al.*, 2019).

As plataformas educacionais impulsionadas pela inteligência artificial têm a capacidade de se ajustar aos estilos individuais de aprendizagem de cada aluno. Essa abordagem personalizada possui o potencial de maximizar os resultados da aprendizagem, assegurando que os estudantes de medicina adquiram habilidades e conhecimentos (SADIKU *et al.*, 2021; NORCINI *et al.*, 2011).

A aprendizagem personalizada e adaptativa foi originalmente proposta e utilizada principalmente em áreas disciplinares como engenharia, idiomas, matemática e ciências, e tende a focar em resultados de aprendizagem que são relativamente simples de medir. Um exemplo de utilização da IA generativa na educação médica é a produção de cenários clínicos simulados que se aproximam bastante dos dados reais dos pacientes. Isso oferece aos alunos uma ampla gama de casos realistas para estudo, reflexões e discussões de patologias específicas, proporcionando aos estudantes a prática e a tomada de decisões clínicas em um ambiente livre de riscos, que não comprometam a integridade e a exposição da identidade do paciente, habituando-os a cenários clínicos desafiadores que serão vivenciados durante sua prática médica (FRID-ADAR *et al.*, 2018).

Nesse contexto de expansão tecnológica, surge a questão de como tais ferramentas podem impactar um dos pilares centrais da formação médica: o feedback educacional. O Feedback é uma estratégia amplamente utilizada no contexto da educação. É descrito como a informação fornecida por professor, colega, autoavaliação ou experiência sobre aspectos do desempenho ou performance de um indivíduo (WATLING; GINSBURG, 2019). É utilizado como um recurso de influência para promover uma melhoria no desempenho acadêmico do aluno, tendo o objetivo de orientar e dar suporte, estimulando a compreensão relacionada ao entendimento e domínio da tarefa (HATTIE; TIMPERLEY, 2007; SHUTE, 2008), permitindo desta forma evidenciar as discrepâncias entre o desempenho real e o pretendido e enfatizar os pontos-chaves que precisam ser trabalhados e otimizados (BOUD; MOLLOY, 2013).

Assim, dialoga diretamente com os princípios da andragogia, proposta por Malcolm Knowles (KNOWLES; HOLTON; SWANSON, 2015), que enfatiza a necessidade de autonomia, relevância prática e aplicação imediata do conhecimento, favorecendo a autorregulação, a reflexão crítica e o engajamento ativo do estudante. Especificamente, o feedback positivo desempenha um papel importante na formação dos estudantes de Medicina, pois reforça comportamentos benéficos e motiva os alunos a aprimorarem suas habilidades (ZEFERINO; DOMINGUES; AMARAL, 2007). As emoções e particularidades precisam ser individualizadas para o educador conseguir transmitir a informação da melhor forma (LIPNEVICH; PANADERO, 2021). O feedback médico adequado fomenta a análise crítica e a autoconfiança, essenciais na construção de profissionais competentes (ZEFERINO; DOMINGUES; AMARAL, 2007). Os discentes terão uma resposta positiva apenas para aquele feedback que é passível de entendimento e de execução, observando o verdadeiro motivo deles terem ocorrido (CARLESS; BOUD, 2018; WISNIEWSKI; ZIERER; HATTIE, 2020).

Embora seja valorizado pelos aprendizes, sua prática ainda é limitada nos ambientes acadêmicos médicos (PRICINOTE; PEREIRA, 2016). A escassez de tempo dos preceptores, as barreiras de comunicação e a resistência dos estudantes de medicina em receber críticas construtivas são alguns desafios que podem ser encontrados (RAMANI *et al.*, 2017). O ensino eficaz não se limita apenas à transmissão de informações e compreensões aos alunos, ou à criação de tarefas e ambientes de aprendizagem (MASTROENI *et al.*, 2022). Também envolve avaliar continuamente a compreensão dos alunos sobre essas informações, para que o próximo passo no ensino possa ser ajustado de acordo com o grau de obtenção de informações pelos alunos (HATTIE; TIMPERLEY, 2007).

Nesse cenário, os professores precisam estar cientes do que cada aluno está pensando e refletindo, para serem aptos a construir experiências significativas aos estudantes. Devem ter conhecimento e compreensão proficientes do conteúdo para fornecê-lo de forma apropriada. Frequentemente, os professores restringem as oportunidades dos alunos de obter feedback sobre seu desempenho, assumindo toda a responsabilidade pelo progresso dos alunos e deixando de explorar as oportunidades de aprendizado próprio (HATTIE; TIMPERLEY, 2007).

Acrescenta-se que ao limitar as oportunidades de diálogo, autorreflexão e construção compartilhada do feedback, o docente afasta-se da perspetiva andragógica, que preconiza ambientes colaborativos, autonomia progressiva e corresponsabilidade do desenvolvimento profissional (KNOWLES; HOLTON; SWANSON, 2015). Entender os estilos dos métodos de ensino, no futuro, permitirão oferecer mais elementos aos tutores na adaptação dos seus estilos às características de aprendizagem do aluno como insumo para fortalecer os perfis docentes e nos processos de qualificação do professor em seu papel de professor tutor (MARTÍNEZ CARRILLO, 2018; BUTLER; WINNE, 1995). Isso proporcionaria aos aprendizes mais flexibilidade, eficácia, adaptação, engajamento e motivação (PENG; MA; SPECTOR,

2019). Foram criados diversos sistemas automatizados para simplificar o processo de entrega de feedback em diversas tarefas educacionais (DAI *et al.*, 2023). À medida que o feedback recebe cada vez mais valorização nos sistemas de aprendizagem, os acadêmicos têm se dedicado ao desenvolvimento de modelos que elucidem como o feedback influencia a aquisição de conhecimentos e quais princípios contribuem para sua concepção eficiente (DAI *et al.*, 2023).

Entretanto, a integração da inteligência artificial (IA) na geração de feedback para estudantes de medicina confronta com alguns desafios significativos. Estudos demonstram que 63,5% dos professores de graduação médica demonstraram apreensão com a desumanização do ensino e 22,2% apontaram o risco de elitização (PRICINOTE; PEREIRA, 2016). Além disso, para a implementação de fato da IA na graduação médica faz-se necessária a implantação de infraestrutura adequada, com o treinamento especializado para os docentes. A falta de familiaridade dos estudantes com a IA e a necessidade de capacitação dos professores são lacunas a serem enfrentadas para uma integração bem-sucedida dessa nova tecnologia (SHUTE, 2008).

Apesar dos desafios, o exposto demonstra que o uso de inteligência artificial (IA) no ensino médico apresenta potencial para viabilizar a oferta de feedback individualizado em larga escala, estratégia reconhecidamente importante para o aprendizado dos estudantes, mas frequentemente limitada por barreiras relacionadas ao dimensionamento das turmas e à disponibilidade docente (REN *et al.*, 2026). Do ponto de vista pessoal e profissional, investigar essa tecnologia torna-se relevante por oferecer subsídios para que educadores possam integrar ferramentas inovadoras ao processo de ensino-aprendizagem, potencializando a qualidade da prática pedagógica sem substituir o papel do professor. Sob a perspectiva social, a utilização de sistemas capazes de fornecer feedback oportuno e consistente pode contribuir para uma formação médica mais eficiente e equitativa, beneficiando estudantes,

instituições de ensino e, indiretamente, a assistência prestada à população por futuros profissionais mais bem preparados. No âmbito acadêmico, observa-se que a literatura ainda carece de estudos que avaliem de forma sistemática a qualidade e a acurácia de feedbacks automatizados gerados por IA no contexto do ensino médico, especialmente em comparação com o feedback humano. Dessa forma, este trabalho objetiva verificar a acurácia e a qualidade de feedbacks automatizados produzidos por IA generativa para estudantes de Medicina, contribuindo para preencher essa lacuna científica e fornecer evidências que subsidiem sua utilização na educação médica.

2 OBJETIVOS

2.1 Objetivo Geral

- Avaliar os feedbacks gerados por ferramentas de IA generativa para respostas de estudantes do curso de Medicina da Universidade Christus.

2.2 Objetivos Específicos

- Analisar a conformidade dos feedbacks fornecidos com as expectativas teóricas do feedback.
- Identificar o grau de comparabilidade de feedbacks fornecidos por IA com o feedback de professores humanos treinados e familiarizados com a oferta de feedback.
- Avaliar a percepção dos estudantes acerca dos feedbacks fornecidos.
- Desenvolver um manual do uso de feedback com IA.

3. REFERENCIAL TEÓRICO

O feedback é amplamente reconhecido como um dos mecanismos pedagógicos mais poderosos para promover aprendizagem significativa, autorregulação e desenvolvimento profissional, especialmente na educação médica, onde o raciocínio clínico e a capacidade de tomar decisões em cenários complexos são habilidades essenciais. Desde a publicação do clássico artigo “The Power of Feedback” por Hattie e Timperley (2007), o conceito evoluiu de uma simples devolutiva corretiva para uma ferramenta estratégica com papel central na orientação do desempenho, na construção da metacognição e no aprimoramento progressivo do estudante (JOHNSON *et al.*, 2020).

Na formação médica, em particular, o feedback tornou-se componente indispensável, pois possibilita que o aluno compreenda não apenas o que errou ou acertou, mas como aprimorar seu processamento clínico, refinar sua abordagem diagnóstica e identificar lacunas conceituais ou práticas (CANTILLON; SARGEANT, 2008).Nas últimas décadas, a literatura em educação médica tem explorado de maneira crescente as dimensões qualitativas do feedback, incluindo sua estrutura, suas funções, sua recepção pelos estudantes e o impacto sobre aprendizagem e desempenho (BOUD; MOLLOY, 2013).

A preocupação com a eficácia do feedback ocorre em um contexto no qual currículos médicos operam sob intensa pressão: grande volume de conteúdo, necessidade de supervisão consistente, crescente número de estudantes por docente e transição para PBL como formato de ensino e aprendizagem ativa. Esses desafios têm alimentado a busca por estratégias inovadoras, incluindo tecnologias emergentes como *learning analytics*, sistemas tutores inteligentes e, recentemente, modelos de linguagem baseados em inteligência artificial (IA) como o ChatGPT (SIMONI *et al.*, 2025).

Segundo Hattie e Timperley (2007), o feedback eficaz responde, de forma estruturada, a três perguntas fundamentais: “Onde estou indo?”, “Como estou indo?” e “Qual o próximo passo?”. Esse modelo distingue três níveis principais: Feedback sobre a tarefa (*task level*): correção do conteúdo, precisão conceitual, acertos e

equívocos. Feedback sobre o processo (*process level*): estratégias cognitivas e analíticas usadas pelo aluno para chegar a uma resposta. Feedback sobre a autorregulação (*self-regulation level*): capacidade de monitorar o próprio aprendizado, identificar erros e utilizar o feedback para melhoria contínua (PHAM *et al.*, 2025).

Essa estrutura evidencia que feedback não é meramente informativo; ele é orientador, formativo e regulador do aprendizado. A literatura contemporânea reforça ainda que o feedback deve ser oportuno, específico, claro e sensível ao contexto (ARCHER, 2010; WATLING; GINSBURG, 2019). Além disso, deve promover segurança psicológica, diálogo e reflexão, pois o modo como os estudantes percebem e utilizam o feedback impacta diretamente sua aprendizagem (SARGEANT *et al.*, 2015).

Diversos estudos mostraram que a qualidade da interação interpessoal entre professor e estudante influencia a recepção do feedback e, portanto, seu impacto (BING-YOU *et al.*, 2017; TELIO *et al.*, 2015). Assim, um bom feedback depende tanto da clareza conceitual quanto da sensibilidade relacional — aspecto que se revelará particularmente relevante quando discutirmos feedback gerado por IA.

3.1 A importância do feedback na educação médica

Na educação médica, o feedback ocupa posição central no processo formativo por múltiplas razões. Primeiramente, constitui elemento essencial para o desenvolvimento do raciocínio clínico, uma vez que a aprendizagem médica ocorre, em grande parte, a partir da análise de erros interpretativos, da identificação de vieses cognitivos e da validação ou reformulação de hipóteses diagnósticas (SCHUWIRTH; VAN DER VLEUTEN, 2011).

Além disso, a prática médica caracteriza-se pela tomada de decisões em contextos de incerteza, exigindo refinamento contínuo do julgamento clínico, sustentado por supervisão qualificada e devolutivas precisas e oportunas. Soma-se a isso o fato de que a educação médica contemporânea se fundamenta predominantemente em metodologias ativas de ensino-aprendizagem, como *problem-based learning*, simulações clínicas, *Objective Structured Clinical Examinations*

(OSCEs) e discussões de caso, estratégias intrinsecamente dependentes de feedback estruturado, significativo e orientador do desempenho discente (TRAN *et al.*, 2025).

Evidências empíricas demonstram que feedback de alta qualidade não apenas impacta positivamente o desempenho imediato dos estudantes, mas também favorece o desenvolvimento da autorregulação, da metacognição e da aprendizagem em longo prazo (MOLLOY; BEARMAN, 2019). Revisões sistemáticas reforçam que feedback eficaz contribui para o desenvolvimento profissional progressivo, amplia a confiança do aprendiz, reduz a recorrência de erros e promove práticas reflexivas fundamentais à formação médica (VELOSKI *et al.*, 2006; WATLING *et al.*, 2021).

Sob a perspectiva freireana, o feedback pode ser compreendido como uma prática dialógica, na qual o educador não se limita à transmissão de julgamentos avaliativos, mas estabelece um processo de comunicação horizontal, crítico e problematizador, capaz de promover a conscientização e a autonomia do aprendiz (FREIRE, 1996).

Nessa lógica, o feedback deixa de ser um ato unidirecional e passa a configurar-se como um espaço de construção compartilhada do conhecimento (ENDE, 1983). De forma convergente, à luz da teoria sociocultural de Lev Vygotsky, o feedback assume papel mediador no processo de aprendizagem, ao atuar como instrumento de apoio dentro da zona de desenvolvimento proximal, possibilitando que o estudante avance de níveis de desempenho assistidos para níveis de desempenho autônomos (VYGOTSKY, 1978). Assim, o feedback qualificado funciona como um mecanismo de mediação pedagógica, ajustado às necessidades do aprendiz e sensível ao contexto de sua prática formativa.

Entretanto, a crescente complexidade do conhecimento médico, associada ao aumento do número de estudantes e às limitações da carga docente, impõe desafios significativos à oferta sistemática de feedback individualizado e de alta qualidade[37]. Esse cenário tem impulsionado o debate acerca do potencial complementar das tecnologias educacionais baseadas em inteligência artificial, especialmente no apoio à geração de feedback estruturado, escalável e alinhado a princípios pedagógicos consolidados.

3.2. Inteligência Artificial no campo educacional

A incorporação da Inteligência Artificial (IA) nos processos educativos tem se expandido nas últimas duas décadas, impulsionada pela evolução de algoritmos capazes de reconhecer padrões complexos, realizar predições e adaptar-se ao comportamento do usuário. Tradicionalmente, a IA na educação concentrou-se em sistemas de tutores inteligentes e plataformas adaptativas, cuja função principal era personalizar trajetórias de aprendizagem e identificar dificuldades específicas dos estudantes (HOLMES *et al.*, 2019). Tais sistemas baseavam-se em modelos estatísticos e regras pré-programadas, apresentando desempenho limitado diante da diversidade e sensibilidade de respostas humanas (VAN DE RIDDER *et al.*, 2008).

Com avanços recentes em redes neurais profundas, a IA atingiu novo patamar de capacidade interpretativa e generativa. Surgiram modelos como GPT, PaLM e LLaMA, capazes de produzir textos coerentes, explicar conteúdos complexos, corrigir raciocínios e simular diálogos instrucionais. Essa mudança marcou a transição de uma IA predominantemente “analítica” para uma IA “generativa” (NING *et al.*, 2025).

Tem-se discutido a emergência de uma IA de natureza reflexiva, capaz de analisar processos, justificar escolhas e oferecer devolutivas metacognitivas, aproximando-se de funções tradicionalmente atribuídas à reflexão pedagógica humana (FLORIDI *et al.*, 2018; SHUM *et al.*, 2023).

3.3. Avaliação da qualidade do feedback

A mensuração da qualidade do feedback apresenta desafios metodológicos relevantes, dado seu caráter interpretativo, situado e dependente da percepção do aprendiz (SHUTE, 2008). Diversos métodos foram propostos, incluindo o uso de rubricas, análise temática, instrumentos validados e esquemas de categorização como o *Teacher Feedback Observation Scheme (TFOS)*. Parte substancial da literatura busca identificar não apenas que características tornam o feedback eficaz, mas também como elas podem ser mensuradas (YU; LIU, 2026).

O estudo de Hattie e Timperley (2007) assinala que feedback tem impacto variável, e que sua eficácia depende mais da qualidade do que da quantidade. Estudos posteriores mostraram que feedback vago, excessivamente elogioso, centrado na pessoa ou meramente descritivo tende a produzir poucos ganhos significativos de aprendizagem (KLUGER; DENISI, 1996; ARCHER, 2010).

Pesquisas em educação médica reforçam que os estudantes valorizam feedback detalhado, específico, estruturado e sensível à sua trajetória de aprendizagem (SARGEANT *et al.*, 2015; WATLING; LINGARD, 2012). Além disso, o caráter relacional do feedback, isto é, o modo como o professor reconhece a identidade e as necessidades do estudante, determina o engajamento com a devolutiva e sua utilização prática. A literatura destaca a importância de distinguir entre feedback (informação fornecida ao aluno) e *feedforward* (orientação sobre os passos futuros), conceito incorporado gradualmente na educação médica contemporânea (MOLLOY; BEARMAN, 2019).

3.4 A avaliação do feedback produzido por inteligência artificial

Com o surgimento dos modelos de linguagem de grande porte, especialmente GPT-3.5, GPT-4 e suas variações, o campo começou a explorar a possibilidade de integrar IA como ferramenta de produção de feedback para estudantes de Medicina (SARGEANT *et al.*, 2015). Estudos recentes têm demonstrado que LLMs conseguem gerar respostas complexas, estruturadas e detalhadas em diversas áreas clínicas, com desempenho comparável ao de estudantes ou até de especialistas em exames teóricos (GILSON *et al.*, 2023; KUNG *et al.*, 2023).

Em pesquisas sobre feedback, a IA demonstrou capacidade de produzir devolutivas altamente detalhadas e extensas, com boa organização textual e explicações aprofundadas de fisiopatologia e diagnóstico (CICEK *et al.*, 2024; GOKKURT YILMAZ *et al.*, 2024). Esse padrão decorre do acesso dos modelos à vasta rede de informações biomédicas presentes no treinamento, o que lhes permite recuperar, recombinar e sintetizar conhecimento clínico em velocidade incomparável à capacidade humana. Assim, enquanto o professor humano apresenta sensibilidade, contextualização e poder de síntese, a IA oferece profundidade, amplitude e consistência de conteúdo.

Entretanto, estudos convergem ao mostrar que, apesar da sofisticação aparente, os feedbacks gerados por IA apresentam limitações importantes: ausência de sensibilidade humana, dificuldade em incorporar nuances relacionais, tendência à polidez artificial e potencial para erros factuais ou “alucinações” (DAI *et al.*, 2024; NADARAJAH *et al.*, 2024). Além disso, pesquisas como as de Watling e Ginsburg (2019) sustentam que o impacto do feedback depende essencialmente da relação entre avaliador e estudante — algo ainda insubstituível pela IA.

Diante da expansão exponencial do conhecimento médico, torna-se inviável que um único docente concentre e atualize, de forma integral, todo o conhecimento disponível na literatura médica contemporânea (DENSEN, 2011). Assim, ferramentas de IA podem atuar como suporte cognitivo avançado, aumentando a escala e a profundidade do feedback, desde que mantida supervisão humana.

3.5. Complementaridade entre feedback humano e IA

O conjunto da literatura mostra que o feedback ideal deve ser específico, detalhado e orientado por evidências (HATTIE; TIMPERLEY, 2007; ARCHER, 2010); contextualizado, relacional e sensível à identidade do estudante (WATLING; LINGARD, 2012; TELIO *et al.*, 2015); e capaz de promover autorregulação e aprendizagem longitudinal (MOLLOY; BEARMAN, 2019).

Entretanto, a incorporação de ferramentas de inteligência artificial nesse processo também suscita importantes questões éticas, como a proteção da privacidade e da confidencialidade dos dados dos estudantes, a transparência dos critérios utilizados pelos algoritmos, a possibilidade de vieses que comprometam a equidade das avaliações e a definição da responsabilidade sobre o conteúdo gerado.

Dessa forma, embora a IA represente uma oportunidade para ampliar o acesso ao feedback individualizado, sua utilização deve ocorrer de maneira crítica, com supervisão docente e alinhada aos princípios éticos da educação médica, garantindo que a tecnologia complemente, e não substitua, o julgamento pedagógico humano.

A IA se destaca essencialmente na primeira dimensão — especificidade e profundidade técnica. Já o professor humano é insubstituível nas dimensões relacional e formativa. Assim, o futuro aponta para modelos híbridos, em que a IA produz um rascunho técnico altamente preciso, e o docente revisa, ajusta e personaliza a devolutiva, promovendo melhor balanceamento entre precisão cognitiva e sensibilidade pedagógica.

4. MATERIAIS E MÉTODOS

4.1 Natureza do estudo

Trata-se de um estudo transversal, observacional de natureza quantitativa, com abordagem analítica.

4.2 Metodologia Proposta

O presente estudo caracteriza-se como um delineamento de validação e observacional, com foco na análise quantitativa da qualidade dos feedbacks gerados por inteligência artificial em comparação aos feedbacks gerados por docentes humanos, bem como na avaliação subsequente de feedbacks gerados por IA pelos estudantes.

4.3 Critérios de Inclusão e Exclusão

Foram incluídas no estudo avaliações teóricas de múltipla escolha previamente aplicadas no curso de Medicina da Universidade Christus, referentes ao primeiro e ao sétimo semestres, elaboradas com base nos conteúdos previstos nas respectivas matrizes curriculares e acompanhadas de gabarito oficial. Também foram considerados elegíveis os feedbacks elaborados a partir das respostas a essas avaliações, tanto por ferramenta de inteligência artificial generativa quanto por docentes humanos, desde que produzidos de acordo com os mesmos critérios metodológicos previamente estabelecidos para análise.

Foram excluídas avaliações ou questões que apresentassem inconsistências de formulação, ausência de gabarito oficial, ambiguidades interpretativas, erros técnico-pedagógicos ou incompatibilidade com os conteúdos programáticos dos semestres selecionados. Da mesma forma, foram excluídos feedbacks incompletos, não padronizados, elaborados em desacordo com os parâmetros metodológicos definidos ou que impossibilitassem a comparação entre os feedbacks gerados pela inteligência artificial e aqueles produzidos pelos docentes.

4.4 População, amostra e local do estudo

A população do estudo foi composta por avaliações teóricas de múltipla escolha do curso de Medicina da Universidade Christus, referentes ao primeiro e ao sétimo semestres, bem como pelos feedbacks produzidos a partir das respostas dos estudantes a essas avaliações. Também integraram o universo da pesquisa os docentes avaliadores vinculados ao curso de Medicina e os estudantes regularmente matriculados nos semestres investigados.

A amostra foi composta por 200 questões de múltipla escolha selecionadas aleatoriamente, sendo 100 referentes ao primeiro semestre, correspondente ao ciclo básico, e 100 referentes ao sétimo semestre, correspondente ao ciclo clínico. Para a etapa de comparação dos feedbacks, participaram 16 docentes avaliadores, vinculados à instituição e com experiência em docência médica, além de três juízes especialistas responsáveis pela análise dos feedbacks gerados pela inteligência artificial e pelos docentes. Na etapa de avaliação da percepção discente, a amostra foi composta por 40 estudantes de Medicina, sendo 20 do primeiro semestre e 20 do sétimo semestre, que avaliaram a qualidade dos feedbacks gerados por inteligência artificial. A pesquisa foi conduzida na cidade de Fortaleza, Ceará, Brasil, nas dependências da Universidade Christus – Campus Parque Ecológico Curso de Medicina, instituição com aproximadamente 2.000 alunos. O período previsto para a execução do estudo compreende os meses de junho de 2024 a janeiro de 2026.

4.4.1. Validação

Procedeu-se à seleção aleatória de 200 questões de múltipla escolha, sendo 100 referentes ao primeiro semestre (ciclo básico) e 100 questões referentes ao sétimo semestre (ciclo clínico), todas elaboradas com base nos conteúdos programados oficialmente previstos nas matrizes curriculares correspondentes aos respectivos períodos do curso de Medicina. A inclusão desses dois semestres fundamenta-se na diferença substancial entre os tipos de competências e demandas cognitivas desenvolvidas nessas etapas do currículo. O tamanho mínimo da amostra

necessário para testar a hipótese nula de concordância κ = 0,3 entre os feedbacks gerados pela IA e pelos avaliadores humanos, contra a hipótese alternativa κ = 0,5, assumindo nível de significância de 5 por cento e poder estatístico de 90 por cento, foi estimado em 173 avaliações. As respostas foram analisadas e julgadas por inteligência artificial e por um painel de 16 docentes avaliadores, vinculados à instituição de ensino, com experiência prévia em docência médica, conforme evidenciado no Flowchart do Estudo (ver apêndice). O sistema de IA utilizado para a geração dos feedbacks foi previamente treinado, calibrado e parametrizado a partir de critérios metodológicos definidos no sistema GPT 4.

4.4.2. Avaliação do feedback pelos estudantes

Na segunda etapa, avaliou-se a qualidade dos feedbacks gerados por inteligência artificial, parametrizada e calibrada com os mesmos critérios metodológicos da etapa anterior, a partir das respostas de 40 estudantes de Medicina, sendo 20 do primeiro e 20 do sétimo semestre, selecionados para representar diferentes níveis de formação. Os feedbacks foram produzidos com base em uma prova de múltipla escolha da própria Instituição, previamente aplicada aos participantes. Os feedbacks foram apresentados aos próprios estudantes, que os avaliaram quanto à sua qualidade por meio da Escala de Avaliação da Qualidade do Feedback Escrito, estruturada em formato de escala Likert, conforme descrito por Lizzio e Wilson (2008).

4.5 Coleta de dados

4.5.1 Parâmetros e procedimentos de mensuração das principais variáveis de Interesse

Na etapa 1, os feedbacks foram classificados segundo critérios descritos pelos modelos Power of Feedback e Teacher Feedback Observation Scheme (TFOS) conforme Hattie & Timperley, 2007 e Thurlings et al, 2012, cujos detalhes encontram-se no tópico Variáveis. Três juízes especialistas, todos com mais de quatro anos de experiência em educação médica, previamente informados e treinados sobre os instrumentos de feedback, realizaram a análise dos feedbacks confeccionados pela

inteligência artificial e pelos 16 professores avaliadores humanos. A escolha do número de juízes fundamentou-se na estratégia metodológica de dupla avaliação independente, na qual dois juízes realizaram inicialmente a análise dos feedbacks de forma cega e autônoma. Nos casos de discordância entre os dois avaliadores primários, o terceiro juiz atuou como avaliador de desempate. As classificações atribuídas pelos juízes especialistas foram posteriormente submetidas à análise estatística comparativa, visando identificar padrões de concordância e discrepâncias entre os feedbacks produzidos pela IA e os feedbacks elaborados pelos professores avaliadores humanos. Na etapa 2, foi selecionada pelos pesquisadores uma prova de múltipla escolha previamente aplicada e respondida por 40 estudantes (20 do primeiro semestre e 20 do sétimo semestre). Esses discentes avaliaram a qualidade dos feedbacks gerados pela IA referentes a essa prova. Essa avaliação foi realizada por meio da Escala de Avaliação da Qualidade do Feedback Escrito, conforme Lizzio & Wilson, 2008 (ver apêndice).

4.6 Variáveis

4.6.1 Avaliação do feedback

Foi utilizada a escala descrita no estudo The Power of Feedback[12], a qual realiza uma análise conceitual do feedback e sintetiza evidências relacionadas ao seu impacto na aprendizagem e nos resultados educacionais. O modelo propõe a identificação das propriedades e das circunstâncias específicas que conferem eficácia ao feedback, abordando o momento mais adequado para sua oferta, bem como os efeitos do feedback positivo e negativo sobre o processo de aprendizagem. Foram desenvolvidos os itens a seguir a serem avaliados pelos juízes:

1 - Feedback correto:
Está correto
Está incorreto
2 - Feedback da tarefa projetado para desencorajar o aluno:
Sim

Não
3 - Feedback de elogios sobre a tarefa:
Sim
Não
4- Complexidade da tarefa:
Muito complexa
Pouco complexa
5- Definição de metas:
Metas difíceis
Metas fáceis, faça o seu melhor
6- Ameaça à autoestima:
Muita ameaça
Pouca ameaça.

4.6.2 Esquema de observação do feedback do professor

O Teacher Feedback Observation Scheme (TFOS)[9] constitui um instrumento destinado à análise do processo de feedback, baseado em dimensões que descrevem suas características, condições e efeitos. Cinco das seis dimensões que descrevem a eficácia do feedback estão incorporadas no TFOS e foram avaliadas pelos 3 juízes:

(1) Direcionamento para a meta versus direcionamento para a pessoa
(2) Específico versus geral
(3) Detalhado versus vago
(4) Corretivo versus não corretivo

(5)Positivo	versus	negativo

4.7 Análise estatística

Os resultados quantitativos categóricos foram apresentados em forma de percentuais e contagens e os numéricos em forma de medidas de tendência central. Foram realizados testes de McNemar e medidas de concordância de Kappa para a comparação dos resultados dos feedbacks humanos e os fornecidos pela IA. Além deles, para variáveis categóricas, foi utilizado o teste de qui-quadrado para verificar associação. Foram considerados significativos valores de p inferiores a 0,05. Os dados obtidos na coleta foram tabulados e analisados pelo software IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp. IBM Corp. Released 2015.

4.8 Avaliação dos riscos e benefícios

A presente pesquisa apresentou um risco mínimo aos envolvidos visto que não existe nenhum procedimento invasivo. Não ocorreu nenhum constrangimento ao responder o questionário. Foi ressaltado ao participante que sua identidade estaria preservada e que em caso de qualquer dúvida quanto a sua participação na pesquisa, os pesquisadores estariam disponíveis para responder quaisquer questionamentos de forma imediata. Os participantes do estudo foram beneficiados com informações sobre melhores feedbacks.

4.9 Aspectos éticos

Este estudo respeitou os preceitos éticos da pesquisa em seres humanos. Foram tomados todos os cuidados no sentido de preservar, em qualquer situação, a identidade e a privacidade dos indivíduos incluídos neste estudo. Não houve nenhum procedimento invasivo. Cada discente e juiz recebeu informações detalhadas sobre os procedimentos, riscos e benefícios, e somente foram incluídos no protocolo após assinatura do Termo de Consentimento Livre e Esclarecido. Não houve nenhum constrangimento ao responder o questionário. O projeto foi submetido ao Comitê de Ética em Pesquisa (CEP) e assumiu perante o mesmo o compromisso de seguir

fielmente os preceitos éticos contidos nas diretrizes e nas normas de pesquisa da resolução 466/2012 do Conselho Nacional de Saúde. O estudo teve início apenas após a aprovação no Comitê de Ética (79594424.6.0000.5049).

5. RESULTADOS

A Tabela 1 apresenta uma análise comparativa abrangente entre a inteligência artificial (IA) e educadores humanos em onze dimensões avaliativas do feedback formativo. Em relação à correção do feedback, a IA foi classificada como fornecedora de feedback correto em 84,1% dos casos (n = 164), enquanto os humanos apresentaram 79,6% (n = 150), embora a concordância entre ambos tenha sido desprezível (κ = 0,025; p = 0,694).

No que se refere à definição de objetivos, a IA e os docentes humanos demonstraram concordância quase perfeita (κ = 0,889; p = 0,000), com ambos identificando objetivos “fáceis” em mais de 89% dos casos (IA n = 174; humanos n = 176). Diferenças significativas na homogeneidade marginal foram observadas para especificidade e nível de detalhamento: a IA forneceu feedback específico em 64,1% dos casos (n = 125), em comparação com 47,2% dos humanos (n = 92; teste de McNemar p = 0,005), e feedback detalhado em 49,2% (n = 96) versus 32,8% dos humanos (n = 64; p = 0,004).

Os valores de Kappa de Cohen para essas duas dimensões foram negativos (κ = -0,343 e κ = -0,175, respectivamente), indicando uma divergência sistemática na forma como as duas fontes aplicam detalhamento e especificidade. Uma concordância moderada foi observada na dimensão de ação corretiva (κ = 0,553; p = 0,000), na qual a IA forneceu feedback corretivo em 90,8% dos casos (n = 177) e os humanos em 89,7% (n = 175).

Para a dimensão de ameaça à autoestima, ambos os agentes predominantemente ofereceram feedback de baixo risco (IA 90,8%, n = 177; humanos 89,2%, n = 174), porém sem concordância entre si (κ = -0,053). Por fim, a valência do

feedback foi majoritariamente positiva tanto para a IA (83,1%, n = 162) quanto para os humanos (86,2%, n = 168), embora a concordância entre avaliadores tenha sido discreta (κ = 0,017; p = 0,812).

Os parâmetros estatísticos completos para todas as variáveis comparativas estão detalhados na Tabela 1.

Tabela 1 : Concordância comparativa e homogeneidade marginal entre o feedback da IA e o feedback humano

Variável de Comparação	**IA Cat. 1 / Humano Cat. 1 (n, %)**	**IA Cat. 1 / Humano Cat. 2 (n, %)**	**IA Cat. 2 / Humano Cat. 1 (n, %)**	**IA Cat. 2 / Humano Cat. 2 (n, %)**	**Totais IA (n, %)**	**Totais humanos (n, %)**	**p(McNemar)**	**K (valor; EP; p)**
Correção (Não × Sim)	8 (17,8%; 25,8%)	23 (15,3%; 74,2%)	37 (82,2%; 22,6%)	127 (84,7%; 77,4%)	Não = 31 (15,9%) Sim = 164 (84,1%)	Não = 45 (23,1%) Sim = 150 (79,6%)	0,092	0,025; EP = 0,072; p = 0,694
Desencoraja o estudante? (Não × Sim)	127 (84,7%; 77,9%)	36 (80,0%; 22,1%)	23 (15,3%; 71,9%)	9 (20,0%; 28,1%)	Não = 163 (83,6%)	Não = 150 (76,9%)	0,117	0,052; EP = 0,074; p = 0,458
Elogio na tarefa (Não × Sim)	144 (84,7%; 86,2%)	23 (92,0%; 13,8%)	26 (15,3%; 92,0%)	2 (8,0%; 7,1%)	Não = 167 (85,6%) Sim = 28 (14,4%)	Não = 170 (87,2%) Sim = 25 (12,8%)	0,775	-0,069; EP = 0,057; p = 0,032
Complexidade da tarefa (Baixa × Alta)	188 (99,5%; 98,9%)	2 (33,3%; 1,1%)	1 (0,5%; 20,0%)	4 (66,7%; 80,0%)	Baixa = 190 (97,4%)	Baixa = 189 (96,9%)	1,000	0,719; EP = 0,155; p = 0,000

					Alta = 5 (2,6%)	Alta = 6 (3,1%)		
Definição de metas (Fácil × Difícil)	173 (98,3%; 99,4%)	1 (5,3%; 0,6%)	3 (1,7%; 14,3%)	18 (94,7%; 85,7%)	Fácil = 174 (89,2%)Difícil = 21 (10,8%)	Fácil = 176 (90,3%) Difícil = 19 (9,7%)	0,625	0,889; EP = 0,055; p = 0,000
Variável de Comparação	IA Cat. 1 / Humano Cat. 1 (n, %)	IA Cat. 1 / Humano Cat. 2 (n, %)	IA Cat. 2 / Humano Cat. 1 (n, %)	IA Cat. 2 / Humano Cat. 2 (n, %)	Totais IA (n, %)	Totais humanos (n, %)	(McNemar)	K (valor; EP; p)
Ameaça à autoestima (Baixa × Alta)	157 (90,2%; 88,7%)	20 (95,2%; 11,3%)	17 (9,8%; 94,4%)	1 (4,8%; 5,6%))	Baixa = 177 (90,8%) Alta = 18 (9,2%)	Baixa = 174 (89,2%) Alta = 21 (10,8%)	0,743	-0,053; EP = 0,055; p = 0,454
Direcionamento (Tarefa × Pessoa)	18 (25,4%; 32,7%)	37 (29,8%; 67,3%)	53 (74,6%; 37,9%)	87 (70,2%; 62,1%)	Tarefa = 140 (71,8%) Pessoa = 55 (28,2%)	Tarefa = 124 (63,6%) Pessoa = 71 (36,4%)	0,113	-0,047; EP = 0,069; p = 0,503
Especificidade (específico × geral)	20 (19,4%; 28,6%)	50 (54,3%; 71,4%)	83 (80,6%; 66,4%)	42 (45,7%; 33,6%)	Específico = 125 (64,1%) Geral = 70 (35,9%)	Específico = 92 (47,2%) Geral = 103 (52,8%)	0,005	-0,343; EP = 0,065; p = 0,000

Detalhamento (detalhado × vago)	58 (44,3%; 58,6%)	41 (64,1%; 41,4%)	73 (55,7%; 76,0%)	23 (35,9%; 24,0%)	Detalhado= 96 (49,2%) Vago= 99 (50,8%)	Detalhado= 64 (32,8%) Vago= 131 (67,2%)	0,004	-0,175; EP = 0,066; p = 0,009
Caráter corretivo (Não × Sim)	11 (55,0%; 61,1%)	7 (4,0%; 38,9%)	9 (45,0%; 5,1%)	168 (96,0%; 94,9%)	Sim = 177 (90,8%) Não = 18 (9,2%)	Sim = 175 (89,7%) Não = 20 (10,3%)	0,804	0,553; EP = 0,103; p = 0,000
Valência (Negativa × Positiva)	5 (18,5%; 15,2%)	28 (16,7%; 84,8%)	22 (81,5%; 13,6%)	140 (83,3%; 86,4%)	Neg = 33 (16,9%) Pos = 162 (83,1%)	Neg = 27 (13,8%) Pos = 168 (86,2%)	0,480	0,017; EP = 0,073; p = 0,812

Fonte: dados da pesquisa, março de 2026.

A análise comparativa da qualidade do feedback entre a inteligência artificial e os avaliadores humanos ao longo dos semestres acadêmicos está apresentada na Tabela 2. No primeiro semestre (ciclo básico), o feedback gerado pela IA mostrou-se significativamente mais específico em comparação à avaliação humana (73,5% versus 51,0%; p = 0,007), além de apresentar maior frequência de orientações detalhadas (65,3% versus 38,8%; p = 0,002).

Embora a especificidade global tenha diferido entre as fontes (64,2% para a IA versus 47,2% para os humanos; p = 0,005), foram observados elevados níveis de concordância positiva na definição de objetivos no primeiro semestre ($\kappa = 1,000$; $p < 0,001$) e na natureza corretiva do feedback no sétimo semestre ($\kappa = 0,795$; $p < 0,001$).

Observou-se, ainda, uma divergência sistemática no sétimo semestre no que se refere à ameaça à autoestima ($\kappa = -0,080$; $p = 0,039$), embora ambos os avaliadores

tenham predominantemente fornecido feedback de baixo potencial ameaçador (90,5% para a IA versus 84,2% para os humanos na categoria “Baixo”).

Os dados completos referentes a todas as variáveis que apresentaram significância estatística encontram-se detalhados na Tabela 2.

Tabela 2 : Indicadores estatísticos comparativos da qualidade do feedback entre a IA e os avaliadores humanos por semestre

Variáveis de comparação	**Semestre**	**Resultado IA (n, %)**	**Resultado humano (n, %)**	**p (McNemar)**	**K (valor; p-valor)**
Complexidade da tarefa	1	Baixa: 95 (96,9%)	Baixa: 95 (96,9%)	1,000	0,656 (p < 0,001)
	7	Baixa: 93 (97,9%)	Baixa: 92 (96,8%)	1,000	0,795 (p < 0,001)
	Total	Baixa: 188 (97,4%)	Baixa: 187 (96,9%)	1,000	0,719 (p < 0,001)
Definição de metas	1	Fácil: 86 (87,8%)	Fácil: 86 (87,8%)	1,000	1,000 (p < 0,001)
	7	Fácil: 86 (90,5%)	Fácil: 88 (92,6%)	0,625	0,727 (p < 0,001)
	Total	Fácil: 172 (89,1%)	Fácil: 174 (90,2%)	0,625	0,880 (p < 0,001)
Específico vs. geral	1	Específico: 72 (73,5%)	Específico: 50 (51,0%)	0,007	-0,278 (p = 0,002)
	7	Específico: 52 (54,7%)	Específico: 41 (43,2%)	0,228	-0,434 (p < 0,001)
	Total	Específico: 124 (64,2%)	Específico: 91 (47,2%)	0,005	-0,336 (p < 0,001)

Detalhado vs. vago	1	Detalhado: 64 (65,3%)	Detalhado: 38 (38,8%)	0,002	-0,299 (p = 0,001)
	Total	Detalhado: 96 (49,7%)	Detalhado: 64 (33,2%)	0,004	-0,183 (p = 0,007)
Caráter corretivo	1	Corretivo: 84 (85,7%)	Corretivo: 81 (82,7%)	0,607	0,426 (p < 0,001)
	7	Corretivo: 92 (96,8%)	Corretivo: 93 (97,9%)	1,000	0,795 (p < 0,001)
	Total	Corretivo: 176 (91,2%)	Corretivo: 174 (90,2%)	0,804	0,510 (p < 0,001)
Direcionamento do feedback	Total	Tarefa: 139 (72,0%)	Tarefa: 123 (63,7%)	0,013	-0,061 (p = 0,389)
Ameaça à autoestima	7	Baixa: 86 (90,5%)	Baixa: 80 (84,2%)	1,000	-0,080 (p = 0,039)

Fonte: dados da pesquisa, março de 2026.

Os resultados da análise comparativa entre a inteligência artificial e os docentes humanos, considerando as categorias de questões fáceis e difíceis, encontram-se sintetizados na Tabela 3. Na categoria de questões fáceis (Índice 1), a IA apresentou frequência significativamente superior de feedback correto em relação aos avaliadores humanos (93,0% versus 76,0%; p = 0,002), além de demonstrar menor propensão a fornecer feedback de caráter desmotivador (11,0% versus 23,0%; p = 0,036).

No conjunto total da amostra, o feedback gerado pela IA foi mais frequentemente classificado como específico (64,1% versus 47,2%; p = 0,005) e detalhado (49,2% versus 32,8%; p = 0,004), embora essas dimensões tenham apresentado valores negativos de Kappa estatisticamente significativos (especificidade total $\kappa = -0,343$; $p < 0,001$; detalhamento total $\kappa = -0,175$; $p = 0,009$), indicando divergência sistemática entre as fontes.

Elevados níveis de concordância estatisticamente significativa foram observados quanto à complexidade da tarefa (κ total = 0,719; p < 0,001), à definição de objetivos (κ total = 0,889; p < 0,001) e à natureza corretiva do feedback (κ total = 0,534; p < 0,001), dimensões nas quais ambas as fontes predominantemente forneceram orientações corretivas (90,8% para a IA versus 89,7% para os humanos).

Os dados completos relativos a todas as variáveis analisadas encontram-se apresentados na Tabela 3.

Tabela 3 : Indicadores estatísticos comparativos da qualidade do feedback entre a IA e os avaliadores humanos

Variáveis de comparação	**Índice de dificuldade**	**Resultado IA (n, %)**	**Resultado humano (n, %)**	**p (McNemar, bicaudal)**	**K (valor; p-valor)**
Feedback correto?	1 (Fácil)	Sim: 93 (93,0%)	Sim: 76 (76,0%)	0,002	0,023 (p = 0,769)
Feedback desencorajador?	1 (Fácil)	Não: 89 (89,0%)	Não: 77 (77,0%)	0,036	0,032 (p = 0,721)
Complexidade da tarefa	1 (Fácil)	Baixa: 99 (99,0%)	Baixa: 98 (98,0%)	1,000	0,662 (p < 0,001)
	2 (Difícil)	Baixa: 91 (95,8%)	Baixa: 91 (95,8%)	1,000	0,739 (p < 0,001)
	Total	Baixa: 190 (97,4%)	Baixa: 189 (96,9%)	1,000	0,719 (p < 0,001)
Definição de metas	1 (Fácil)	Fácil: 87 (87,0%)	Fácil: 88 (88,0%)	1,000	0,954 (p < 0,001)
	2 (Difícil)	Fácil: 87 (91,6%)	Fácil: 88 (92,6%)	1,000	0,783 (p < 0,001)
	Total	Fácil: 174 (89,2%)	Fácil: 176 (90,3%)	0,625	0,889 (p < 0,001)
Variáveis de comparação	**Índice de dificuldade**	**Resultado IA (n, %)**	**Resultado humano (n, %)**	**p (McNemar, bicaudal)**	**K (valor; p-valor)**
Específico vs. Geral	1 (Fácil)	Específico: 66 (66,0%)	Específico: 44 (44,0%)	0,01	-0,310 (p = 0,001)
	Total	Específico: 125 (64,1%)	Específico: 92 (47,2%)	0,005	-0,343 (p < 0,001)
Detalhado vs. Vago	2 (Difícil)	Detalhado: 46 (48,4%)	Detalhado: 35 (36,8%)	0,185	-0,210 (p = 0,035)

	Total	Detalhado: 96 (49,2%)	Detalhado: 64 (32,8%)	0,004	-0,175 (p = 0,009)
Natureza Corretiva	1 (Fácil)	Corretivo: 93 (93,0%)	Corretivo: 92 (92,0%)	1	0,496 (p < 0,001)
	2 (Difícil)	Corretivo: 84 (88,4%)	Corretivo: 83 (87,4%)	1,000	0,555 (p < 0,001)
	Total	Corretivo: 177 (90,8%)	Corretivo: 175 (89,7%)	0,804	0,534 (p < 0,001)

Fonte: dados da pesquisa, março de 2026.

A Figura 1 apresenta as avaliações dos participantes quanto à qualidade do feedback em 13 itens de análise, evidenciando que a resposta mais frequente para todos os itens foi “concordo totalmente”. As maiores proporções de concordância plena foram observadas nos itens legibilidade (96,9%), alinhamento com os critérios de avaliação (93,8%) e ausência de contradições (87,5%).

Também foram identificadas elevadas proporções de “concordo totalmente” nos itens clareza e compreensibilidade e na percepção de que o feedback contribuiu para a melhoria do desempenho, ambos com 84,4%, bem como no tom respeitoso e motivador (81,2%), no reconhecimento do esforço (78,1%) e na valorização dos aspectos positivos da resposta (78,1%).

Os itens relacionados à reflexão mais aprofundada, à aplicabilidade em avaliações futuras, à clareza dos critérios esperados, à oferta de estratégias de melhoria e à compreensão dos pontos que necessitavam de aprimoramento apresentaram proporções ligeiramente inferiores de “concordo totalmente”, variando entre 68,8% e 71,9%, com pequenas distribuições entre “concordo fortemente”, “concordo parcialmente” e, em alguns itens, “neutro/indiferente”.

De modo geral, a Figura 1 revela um padrão consistente de avaliações favoráveis em todas as dimensões da qualidade do feedback.

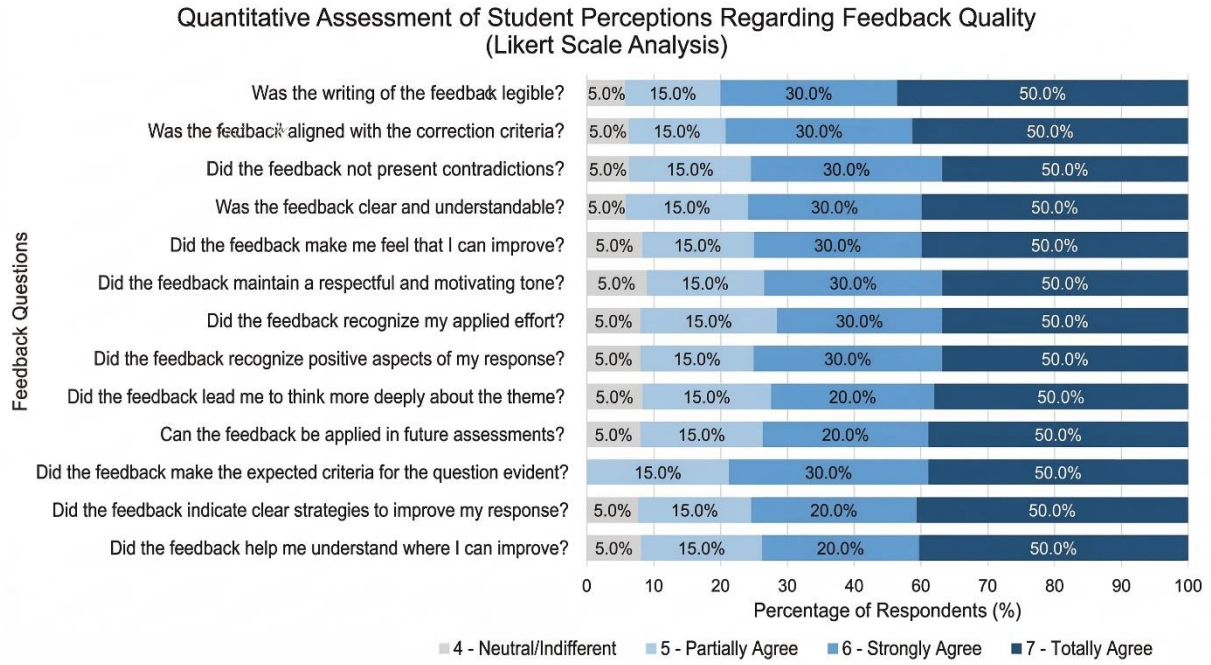


Figura 1. Distribuição das avaliações dos participantes quanto à qualidade do feedback em 13 itens de análise. As barras representam a porcentagem de respostas para cada item em uma escala de concordância de quatro pontos: 4 = neutro/indiferente, 5 = concordo parcialmente, 6 = concordo fortemente e 7 = concordo totalmente.

6. DISCUSSÃO

Este estudo demonstra que, embora a inteligência artificial generativa (IA) possua notável capacidade de produzir feedback formativo tecnicamente acurado, altamente específico e amplamente detalhado, ela apresenta divergências pedagógicas sistemáticas quando comparada aos educadores humanos. Os dados revelam um fenômeno paradoxal: a IA supera significativamente os docentes na geração de detalhamento estrutural e de explicações amplas, baseadas em regras, especialmente nas etapas iniciais do ciclo básico. Contudo, os educadores humanos mantêm uma vantagem crítica, ainda não replicável, no que se refere à especificidade pedagógica, definida como a habilidade refinada de identificar lacunas cognitivas subjacentes e de contextualizar o feedback corretivo em cenários de raciocínio clínico complexo.

Nesse sentido, os coeficientes negativos de Kappa de Cohen observados nas dimensões de especificidade e detalhamento indicam que, embora o feedback da IA e o humano possam compartilhar uma superfície linguística semelhante e elevadas taxas de correção factual, ambos desempenham, na atualidade, funções pedagógicas fundamentalmente distintas no ecossistema da educação médica.

Os testes de McNemar evidenciaram discordâncias sistemáticas entre IA e avaliadores humanos nas categorias “detalhado versus vago” e “específico versus geral”, com coeficientes de Kappa negativos, indicando divergência superior à esperada pelo acaso. Tal divergência é particularmente relevante, uma vez que detalhamento e especificidade constituem pilares essenciais do feedback de alta potência, conforme sintetizado na base de dados de 74 meta-análises de John Hattie e Helen Timperley, que reúne mais de 7.000 estudos sobre intervenções pedagógicas e desempenho acadêmico.

Os resultados sugerem que o ChatGPT apresenta elevada capacidade de produzir explicações extensas, com grande riqueza lexical e de conteúdo, em consonância com estudos que demonstram o potencial dos modelos de linguagem de grande escala na elaboração detalhada de conteúdos, como evidenciado por Dai et al. e Kung et al., nos quais o ChatGPT apresentou desempenho próximo ou

equivalente ao limiar de aprovação em 305 questões de exames de residência médica nos Estados Unidos.

No presente estudo, a IA apresentou melhor desempenho no que se refere ao nível de detalhamento, especialmente no primeiro semestre, uma vez que essa etapa do curso exige predominantemente reconhecimento de padrões, descrição conceitual e organização de conteúdos factuais, com menor demanda por inferências clínicas complexas. Ademais, os docentes humanos não conseguem reter integralmente o volume de conhecimento mobilizado pela IA para aplicá-lo na elaboração do feedback requerido. Conforme argumentado por Herrmann-Werner A et al., o desempenho da IA mostrou-se inferior, com rendimento de 78%, nos níveis da Taxonomia de Bloom que exigem aplicação de conceitos a novas situações, a partir da análise da acurácia de 307 questões originais de múltipla escolha aplicadas a estudantes de medicina na Alemanha.

Entretanto, a profundidade de especificidade e detalhamento observada nos resultados produzidos pela IA nem sempre corresponde à especificidade pedagógica, entendida como a capacidade de identificar a lacuna cognitiva do estudante. Tal limitação torna-se evidente em situações que demandam julgamento pedagógico e análise aprofundada do raciocínio clínico discente, como demonstrado por Seßler, Bewersdorff, Nerdel et al., em estudo que analisou 27 feedbacks, sendo 9 gerados por modelos de linguagem de grande escala (LLMs), 9 por docentes e 9 por especialistas em educação. A avaliação foi conduzida por quatro examinadores cegos com expertise formal na área, considerando dimensões como relevância pedagógica, adequação ao erro cometido pelo estudante e orientação para etapas subsequentes. Esses achados corroboram os resultados do presente estudo, no qual se observou que o feedback fornecido pela IA foi considerado mais corretivo apenas nos semestres iniciais.

Por outro lado, os docentes humanos demonstram maior capacidade de ajustar suas respostas de acordo com o contexto educacional e o tipo de raciocínio envolvido, dimensões frequentemente ausentes no feedback automatizado. Tal fato contribui para explicar por que os coeficientes de Kappa se aproximaram de zero ou assumiram valores negativos: embora semelhantes na superfície linguística, o feedback humano e o da IA não desempenham a mesma função pedagógica.

É amplamente reconhecido que fatores internos, como confiança, medo, autoimagem, experiência e processos de raciocínio cognitivo-emocional, modulam a receptividade ao feedback, conforme demonstrado por Kevin W. Eva et al., em estudo no qual 134 participantes de cinco países avaliaram suas experiências com feedback e os motivos que os levaram a aceitá-lo ou rejeitá-lo, a partir da narrativa de situações reais em que receberam feedback, bem como das percepções e sentimentos associados a esses momentos.

A sensibilidade humana na interpretação do raciocínio subjacente aos erros em questões de múltipla escolha pode ser evidenciada no presente estudo, o qual demonstrou que 97% dos feedbacks humanos ($p < 0,005$) referentes ao desempenho de estudantes do sétimo semestre foram classificados como complexos pelos avaliadores. Tal achado evidencia a capacidade sutil do avaliador humano de identificar aplicações inadequadas do raciocínio clínico ou interpretações equivocadas do enunciado por parte do discente. Essa habilidade interpretativa não foi reproduzida pela IA no presente estudo, cujas respostas, embora tecnicamente corretas, tenderam a não abordar a origem real do erro, corroborando os achados de Ali M et al., nos quais 108 estudantes do segundo ano de medicina julgaram o tutor humano superior, especialmente em aspectos relacionados ao julgamento, à capacidade de orientar a ação e ao foco na lacuna específica do estudante, ao responderem questionários online acerca da clareza, utilidade e completude do feedback.

Em cenários clínicos de semestres mais avançados, nos quais o raciocínio incorreto pode emergir de complexidades cognitivas, essa sensibilidade humana torna-se essencial para a efetividade do feedback. Tal fenômeno foi igualmente observado neste estudo, no qual os docentes ofereceram explicações contextualizadas, ajustadas ao nível de progressão curricular e às dificuldades típicas daquele momento da formação. No sétimo semestre, os resultados evidenciaram superioridade do feedback corretivo humano. Esse achado é convergente com o estudo de Çiçek *et al*., ao indicar que o feedback gerado por IA pode carecer das nuances necessárias para casos clínicos mais complexos e interpretativos.

A análise demonstrou que, no que se refere às categorias de positividade, elogio e ameaça à autoestima, ambos os grupos apresentaram desempenho semelhante, sem diferenças estatisticamente significativas, embora estudos prévios

indiquem que feedback excessivamente positivo ou excessivamente brando pode prejudicar a autorregulação e reduzir o engajamento, conforme observado por Kearney *et al*., especialmente entre estudantes em fases iniciais da formação.

Diversas limitações relevantes devem ser consideradas na interpretação dos achados deste estudo. Em primeiro lugar, trata-se de uma investigação de delineamento observacional e transversal, conduzida em uma única instituição de ensino médico, o que restringe a generalização imediata dos resultados para diferentes currículos, culturas pedagógicas e contextos linguísticos. Em segundo lugar, a metodologia empregada avaliou estritamente a qualidade formal e estrutural do feedback escrito, à luz de modelos teóricos consolidados (TFOS e Power of Feedback), por meio de julgamento de especialistas. Todavia, elevada qualidade estrutural não implica, necessariamente, eficácia pedagógica funcional, uma vez que o estudo não avaliou, de forma longitudinal, desfechos educacionais subsequentes, como mudanças comportamentais sustentadas, integração do feedback ao raciocínio clínico posterior ou desempenho acadêmico final.

Em terceiro lugar, a análise encontra-se intrinsecamente condicionada à efemeridade algorítmica; embora o modelo de IA generativa utilizado (GPT-4) tenha sido cuidadosamente parametrizado por meio de estratégias rigorosas de prompting, a evolução rápida, opaca e contínua dos modelos de linguagem de grande escala implica que versões futuras possam apresentar perfis pedagógicos e linguísticos substancialmente distintos.

Por fim, embora esta etapa de validação se baseie exclusivamente na avaliação robusta de docentes especialistas, o impacto pedagógico real é dependente do usuário. Investigações futuras devem incorporar avaliações diretas, em larga escala, por parte dos estudantes, a fim de elucidar como estes percebem, interpretam e operacionalizam o feedback automatizado em contextos de aprendizagem de alta exigência.

Os resultados indicam que, embora a IA generativa seja capaz de produzir feedback detalhado e tecnicamente bem estruturado, ela não reproduz os padrões de sutileza interpessoal e refinamento pedagógico característicos do feedback humano. Por outro lado, a convergência observada em itens sem significância estatística

sugere que a IA pode atuar como ferramenta de apoio, especialmente na ampliação do acesso ao feedback, sobretudo em contextos marcados por desproporção entre o número de docentes disponíveis e o contingente discente.

Os resultados deste estudo não permitem afirmar uma superioridade global da inteligência artificial em relação ao feedback elaborado por docentes humanos. Embora a IA tenha demonstrado melhor desempenho em dimensões técnicas específicas, como especificidade e detalhamento das devolutivas, tais características, isoladamente, não garantem maior qualidade pedagógica ou impacto formativo. O feedback docente permanece insubstituível em aspectos relacionais, contextuais e interpretativos, sendo capaz de considerar nuances do raciocínio clínico, das necessidades individuais do estudante e do contexto educacional. Assim, os achados sugerem que a maior contribuição da IA reside em seu potencial como ferramenta complementar ao trabalho do professor, ampliando a escalabilidade e a padronização do feedback, sem substituir o julgamento pedagógico humano.

Pesquisas futuras devem considerar o desenvolvimento de prompts alinhados a modelos pedagógicos específicos, previamente treinados e validados, bem como a diferenciação de níveis de feedback e a implementação de modelos híbridos de coautoria entre humanos e inteligência artificial. Nessa perspectiva, o aprimoramento dessas tecnologias deve buscar integrar a precisão técnica proporcionada pela IA com a sensibilidade contextual, ética e formativa do docente, potencializando a qualidade do processo de ensino-aprendizagem na educação médica.

7. **CONCLUSÃO**

O presente estudo permite reconhecer que a inteligência artificial generativa vem assumindo, de forma cada vez mais concreta, um papel relevante no cenário da educação médica, sobretudo no que diz respeito à ampliação do acesso ao feedback formativo. Observou-se que, em diferentes situações, a IA foi capaz de produzir devolutivas corretas do ponto de vista técnico, além de mais específicas e detalhadas quando comparadas àquelas elaboradas por docentes, especialmente no que tange conteúdos de natureza mais fundamental e de menor complexidade cognitiva. Esse desempenho sugere um potencial importante como instrumento de apoio ao processo de ensino, particularmente em contextos nos quais o tempo do professor é limitado e as demandas pedagógicas se impõem de maneira mais significativa.

Os resultados também evidenciam que a qualidade do feedback não se restringe à sua correção ou nível de detalhamento. A atuação do docente humano mostrou-se superior na capacidade de interpretar nuances do raciocínio clínico, identificar lacunas cognitivas e oferecer direcionamentos ajustados ao nível de desenvolvimento do estudante, especialmente no ciclo clínico. Essa dimensão, mais sensível e contextual, permanece como um diferencial essencial da mediação humana no processo educativo, destacando que o feedback eficaz envolve não apenas informação, mas também julgamento pedagógico e compreensão do aprendiz em sua singularidade.

Entretanto, esses achados devem ser interpretados à luz de algumas limitações. Trata-se de um estudo de delineamento transversal, realizado em uma única instituição de ensino, o que pode limitar a generalização dos resultados para outros contextos educacionais. Além disso, a pesquisa avaliou a qualidade e a acurácia dos feedbacks em um momento específico, sem investigar longitudinalmente seu impacto sobre a aprendizagem, a retenção do conhecimento ou o desenvolvimento do raciocínio clínico dos estudantes. Dessa forma, embora os resultados ofereçam evidências relevantes sobre o potencial da inteligência artificial na geração de feedbacks educacionais, estudos multicêntricos e de acompanhamento

longitudinal são necessários para confirmar sua efetividade e aplicabilidade em diferentes cenários da educação médica.

Dessa forma, os dados apontam para uma relação de complementaridade, e não de substituição, entre inteligência artificial e educadores. A incorporação da IA no ensino médico deve ser conduzida de maneira crítica, estratégica e contextualizada, reconhecendo suas potencialidades, mas também seus limites. Seu uso pode ampliar o acesso a feedbacks pedagogicamente consistentes.

Por fim, este estudo contribui para o avanço das discussões sobre o papel da inteligência artificial na educação, indicando caminhos possíveis para sua integração responsável. Investigações futuras são necessárias para aprofundar a compreensão do seu impacto longitudinal na aprendizagem e para desenvolver modelos híbridos que preservem a dimensão humana do ensino, ao mesmo tempo em que se beneficiam das inovações tecnológicas disponíveis

8. ARTIGO ENVIADO PARA PUBLICAÇÃO

Evaluating the Pedagogical Quality of Automated Generative AI Feedback Versus Human Faculty in Medical Education: An Observational Study

Marcelo Lima Gonzaga, MD[1], Hermano Alexandre Lima Rocha, MD, PhD[1,2]

Marcos Kubrusly, MD, PhD[1],

Damile Pinheiro Machado1, Isadora Néri Viana[1],

Ana Beatriz Teófilo Macedo dos Santos[1],

[1] Unichristus University Center, R. João Adolfo Gurgel, 133 - Cocó, Fortaleza - CE, 60190-180, Brazil;

[2] Federal University of Ceará, Community Health Department, R. Papi Júnior, 1223 - Rodolfo Teófilo, Fortaleza - CE, 60430-235, Brazil.

Address correspondence to: Hermano AL Rocha, PhD, Rua Papi Júnior, 1223, 5th. Floor, Fortaleza, Ceará, Brazil. Email: hermano@ufc.br

Short Title

Human-AI Collaboration in Medical Education

9. PRODUTO TÉCNICO PRINCIPAL

REFERÊNCIAS

BLACK, Paul; WILIAM, Dylan. Assessment and classroom learning. **Assessment *in* Education: principles, policy & practice**, v. 5, n. 1, p. 7-74, 1998.

BOUD, David; MOLLOY, Elizabeth. Rethinking models of feedback for learning: the challenge of design. **Assessment & Evaluation *in* higher education**, v. 38, n. 6, p. 698-712, 2013.

BUTLER, Deborah L.; WINNE, Philip H. Feedback and self-regulated learning: A theoretical synthesis. **Review of educational research**, v. 65, n. 3, p. 245-281, 1995.

CANTILLON, Peter; SARGEANT, Joan. Giving feedback in clinical settings. **BMJ**, v. 337, 2008.

CARLESS, David; BOUD, David. The development of student feedback literacy: Enabling uptake of feedback. **Assessment & evaluation *in* higher education**, v. 43, n. 8, p. 1315-1325, 2018.

CHAO, M. A. Generated Artificial General Intelligence: The Philosophical Principle of Artificial General Intelligence and Give an Example. **J Biol Today's World**, v. 11, n. 5, p. 001-010, 2022.

DAI, Wei *et al*. Can large language models provide feedback to students? A case study on ChatGPT. In: **2023 IEEE international conference on advanced learning technologies (ICALT)**. IEEE, 2023. p. 323-325.

DENSEN, Peter. Challenges and opportunities facing medical education. **Transactions of the American clinical and climatological association**, v. 122, p. 48, 2011.

ENDE, Jack. Feedback in clinical medical education. **Jama**, v. 250, n. 6, p. 777-781, 1983.

FRID-ADAR, Maayan *et al*. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. **Neurocomputing**, v. 321, p. 321-331, 2018.

GAŠEVIĆ, Dragan; SIEMENS, George; SADIQ, Shazia. Empowering learners for the age of artificial intelligence. **Computers and education: artificial intelligence**, v. 4, p. 100130, 2023.

GOODFELLOW, Ian J. *et al*. Generative adversarial nets. **Advances in neural information processing systems**, v. 27, 2014.

HATTIE, John; TIMPERLEY, Helen. The power of feedback. **Review of educational research**, v. 77, n. 1, p. 81-112, 2007.

JOHNSON, Christina Elizabeth; WEERASURIA, Mihiri P.; KEATING, Jennifer L. Effect of face-to-face verbal feedback compared with no or alternative feedback on the objective workplace task performance of health professionals: a systematic review and meta-analysis. **BMJ open**, v. 10, n. 3, p. e030672, 2020.

KNOWLES, Malcolm S.; HOLTON, E. F. I.; SWANSON, Richard A. **The adult learner: The definitive classic in adult education and human**. Florence: Taylor and Francis, v. 265, 2015.

LI, Bohu *et al*. New generation artificial intelligence-driven intelligent manufacturing (NGAIIM). *In*: **2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud &**

Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, 2018. p. 1864-1869. Disponível em: https://doi.org/10.1109/smartworld.2018.00313.

LI, Jinlei *et al*. Application of AI-Generated Content *in* Medical Education: Systematic Review of the Impact on Critical Thinking Abilities of Medical Students. **JMIR Medical Education**, v. 12, n. 1, p. e79939, 2026.

LIPNEVICH, Anastasiya A.; PANADERO, Ernesto. A review of feedback models and theories: Descriptions, definitions, and conclusions. In: **Frontiers in education**. Frontiers Media SA, 2021. p. 720195.

LODGE, Jason M.; THOMPSON, Kate; CORRIN, Linda. Mapping out a research agenda for generative artificial intelligence *in* tertiary education. **Australasian Journal of Educational Technology**, v. 39, n. 1, p. 1-8, 2023.

MARTÍNEZ CARRILLO, Esmeralda. **Una aproximación a los estilos de retroalimentación en tutoría de trabajo de grado en educación superior**. 2018.

MASTROENI, M. *et al*. Artificial intelligence *in* medical education: a systematic review. **BMC Medical Education**, v. 22, 2022.

MIAO, Fengchun; HOLMES, Wayne. **Artificial intelligence and education. Guidance for policy-makers**. 2021.

NING, Yilin *et al*. How can artificial intelligence transform the training of medical students and physicians? **The Lancet Digital Health**, v. 7, n. 10, 2025.

NORCINI, John *et al*. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. **Medical teacher**, v. 33, n. 3, p. 206-214, 2011.

PARDO, Abelardo *et al*. Using learning analytics to scale the provision of personalised feedback. **British journal of educational technology**, v. 50, n. 1, p. 128-138, 2019.

PENG, Hongchao; MA, Shanshan; SPECTOR, Jonathan Michael. Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment. **Smart learning environments**, v. 6, n. 1, p. 1-14, 2019.

PHAM, Thai Duong *et al*. The impact of generative AI on health professional education: a systematic review *in* the context of student learning. **Medical Education**, v. 59, n. 12, p. 1280-1289, 2025.

PRICINOTE, Sílvia Cristina Marques Nunes; PEREIRA, Edna Regina Silva. Percepção de Discentes de Medicina sobre o Feedback no Ambiente de Aprendizagem. **Revista Brasileira de Educação Médica**, v. 40, n. 3, p. 470-480, 2016.

RAMANI, Subha *et al*. “It's just not the culture”: a qualitative study exploring residents' perceptions of the impact of institutional culture on feedback. **Teaching and learning in medicine**, v. 29, n. 2, p. 153-161, 2017.

REN, Y. *et al*. Application of Artificial Intelligence in Medical Education: A Systematic and Narrative Review of Pedagogical Potential and Ethical Implications. **Adv Med Educ Pract**, v. 17, p. 567190, 2026. Disponível em: https://doi.org/10.2147/AMEP.S567.

SADIKU, Matthew N. O. *et al*. Artificial Intelligence in Education. **International Journal of Scientific Advances (IJSCIA)**, v. 2, n. 1, p. 5-11, jan./fev. 2021. Disponível em: https://www.ijscia.com/wp-content/uploads/2021/01/Volume2-Issue1-Jan-Fab-No.34-5-11.pdf.

SADLER, D. Royce. Formative assessment and the design of instructional systems. **Instructional science**, v. 18, n. 2, p. 119-144, 1989.

SARGEANT, J. *et al*. Facilitated reflective performance feedback... **Academic Medicine**, v. 90, n. 12, p. 1698–1706, 2015.

SAVAGE, Thomas Robert. Artificial intelligence *in* medical education. **Academic Medicine**, v. 96, n. 9, p. 1229-1230, 2021.

SHUTE, Valerie J. Focus on formative feedback. **Review of educational research**, v. 78, n. 1, p. 153-189, 2008.

SIMONI, Jennifer *et al*. Artificial intelligence *in* undergraduate medical education: an updated scoping review. **BMC Medical Education**, v. 25, n. 1, p. 1609, 2025.

THURLINGS, Marieke *et al*. Development of the Teacher Feedback Observation Scheme: Evaluating the quality of feedback *in* peer groups. **Journal of Education for Teaching**, v. 38, n. 2, p. 193-208, 2012.

TRAN, Cecilia *et al*. Perceptions and use of generative artificial intelligence *in* medical students: a multicenter survey. **Journal of Medical Education and Curricular Development**, v. 12, p. 23821205251391969, 2025.

VAN DE RIDDER, J. M. Monica *et al*. What is feedback in clinical education? **Medical education**, v. 42, n. 2, p. 189-197, 2008.

WANG, Shan *et al*. Artificial intelligence *in* education: A systematic literature review. **Expert systems with applications**, v. 252, p. 124167, 2024.

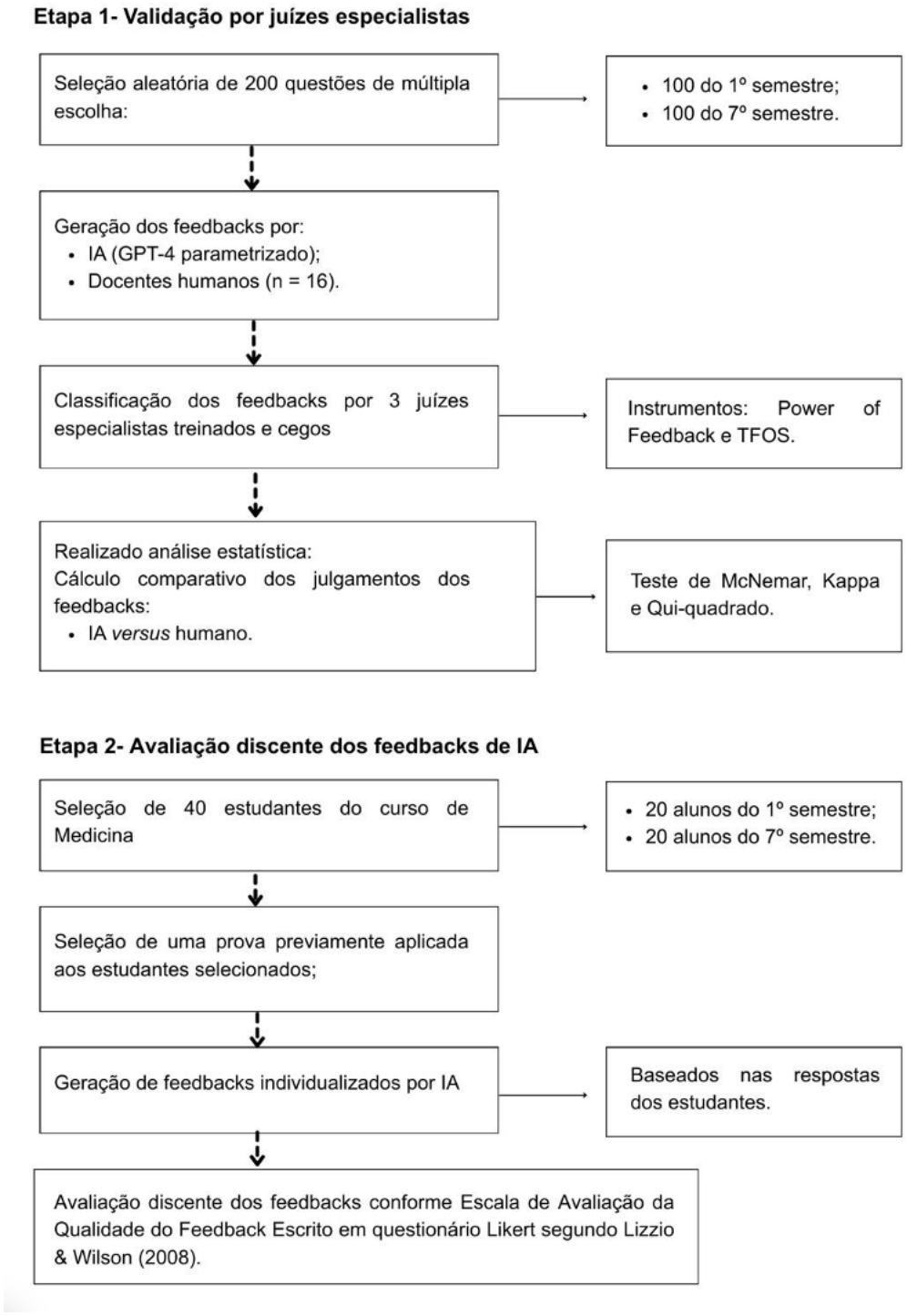
WATLING, Christopher J.; GINSBURG, Shiphra. Assessment, feedback and the alchemy of learning. **Medical education**, v. 53, n. 1, p. 76-85, 2019.

WISNIEWSKI, Benedikt; ZIERER, Klaus; HATTIE, John. The power of feedback revisited: A meta-analysis of educational feedback research. **Frontiers in psychology**, v. 10, p. 487662, 2020.

YU, Sha; LIU, Jin. Mapping the landscape of AI-assisted formative feedback in medical education: A bibliometric analysis. **Medicine**, v. 105, n. 5, p. e47489, 2026.

ZEFERINO, Angélica Maria Bicudo; DOMINGUES, Rosângela Curvo Leite; AMARAL, Eliana. Feedback como estratégia de aprendizado no ensino médico. **Revista Brasileira de Educação Médica**, v. 31, p. 176-179, 2007.

APÊNDICE A – FLOWCHART DO DESENHO DO ESTUDO



ANEXO A - TERMO DE CONSENTIMENTO DE LIVRE ESCLARECIDO (TCLE)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Unichristus
Centro Universitário Christus

Termo de Autorização para Uso de Prova

Eu, _____, aluno(a) do curso de Medicina, **autorizo voluntariamente a utilização da minha prova e das minhas respostas referentes à atividade** para fins acadêmicos e de pesquisa.

Declaro estar ciente de que:

1. Minhas respostas serão utilizadas exclusivamente para a elaboração de **feedbacks educacionais** no âmbito do projeto de pesquisa;
2. Após a geração dos feedbacks, poderei ser convidado(a) a **avaliar a clareza, a utilidade e a qualidade** desses feedbacks;
3. A participação é **voluntária**, não acarretando qualquer benefício ou prejuízo acadêmico, e posso desistir a qualquer momento, sem necessidade de justificativa.
4. As informações coletadas serão tratadas com **confidencialidade**, garantindo o anonimato dos participantes em todas as etapas da pesquisa, análises e publicações.

Confirmo que fui devidamente esclarecido(a) sobre os objetivos do estudo, a forma de participação e o uso acadêmico dos dados fornecidos.

Local e data: _____

Assinatura do(a) aluno(a): _____

ANEXO B - AUTORIZAÇÃO DA INSTUIÇÃO

Unichristus
Centro Universitário Christus

CARTA DE ANUÊNCIA INSTITUCIONAL

Declaro, em nome do Centro Universitário Christus – UNICHRISTUS, que estou ciente da parceria no projeto de pesquisa denominado: *"Aplicativo para conhecimento de princípios ativos de enxaguatórios bucais e dentifrícios presentes no mercado"*, do Mestrado Profissional em Ensino em Saúde e Tecnologias Educacionais, cujo orientador é o Professor Dr. Danilo Lopes Ferreira Lima do Curso de Odontologia desse Centro Universitário. Alego que conheço as responsabilidades desta Instituição como coparticipante no presente projeto de pesquisa, contribuindo com a estrutura física, ficando os insumos e os materiais de consumo sob a responsabilidade do Pesquisador, e que, nesses termos, concordo com esta parceria.

Declaro, ainda, que conheço as resoluções éticas brasileiras e cumpro com todas elas, em especial, a Resolução CNS nº 196/96. Estou ciente de que o referido projeto de pesquisa está sendo submetido e somente poderá ser iniciado após aprovação pelo Comitê de Ética em Pesquisa.

Fortaleza, 5 de março de 2020

Danielle Barbosa
Danielle Pinto Bardawil Barbosa
Supervisora de Campus - UNICHRISTUS

Danielle Barbosa
Supervisão de Campus
Centro Universitário Christus
UNICHRISTUS

Campus Benfica
Rua Princesa Isabel, 1920
60015-061 - Fortaleza-CE
Fone: 85.3214.8770 | 3214.8771

Campus Dionisio Torres
Rua Israel Bezerra, 630
60135-460 - Fortaleza-CE
Fone: 85.3257.2020 | Fax: 85.3277.1762

Campus D. Luis
Av. Dom Luis, 911
60160-230 - Fortaleza-CE
Fone: 85.3457.5300 | Fax: 85.3457.5374

Campus Parque Ecológico
Rua João Adolfo Gurgel, 133
60192-345 - Fortaleza-CE
Fone: 85.3265.8100 | Fax: 85.3265.8110

ANEXO C- COMPROVANTE DE ENVIO DO ARTIGO CIENTÍFICO

Evaluating the Pedagogical Quality of Automated Generative AI Feedback Versus Human Faculty in Medical Education: An Observational Study

Marcelo Lima Gonzaga, MD[1], Hermano Alexandre Lima Rocha, MD, PhD[1,2]

Marcos Kubrusly, MD, PhD[1],

Damile Pinheiro Machado1, Isadora Néri Viana[1],

Ana Beatriz Teófilo Macedo dos Santos[1],

[1] Unichristus University Center, R. João Adolfo Gurgel, 133 - Cocó, Fortaleza - CE, 60190-180, Brazil;

[2] Federal University of Ceará, Community Health Department, R. Papi Júnior, 1223 - Rodolfo Teófilo, Fortaleza - CE, 60430-235, Brazil.

Address correspondence to: Hermano AL Rocha, PhD, Rua Papi Júnior, 1223, 5th. Floor, Fortaleza, Ceará, Brazil. Email: hermano@ufc.br

Short Title

Human-AI Collaboration in Medical Education

ANEXO D- PARECER DO COMITÊ DE ÉTICA

CENTRO UNIVERSITÁRIO
CHRISTUS - UNICHRISTUS

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: AVALIAÇÃO DO FEEDBACK FORNECIDO POR FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA NO ENSINO MÉDICO: ESTUDO DE INTERVENÇÃO RANDOMIZADO
Pesquisador: Hermano Alexandre Lima Rocha
Área Temática:
Versão: 1
CAAE: 79594424.6.0000.5049
Instituição Proponente: Instituto para o Desenvolvimento da Educação Ltda-IPADE/Faculdade
Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 6.857.559

Apresentação do Projeto:
Os modelos de inteligência artificial generativa têm recebido bastante destaque devido à sua habilidade em produzir resultados realistas e de alta
qualidade em diversos campos, incluindo geração de imagens e síntese de texto1. Esses modelos são concebidos para compreender e reproduzir
padrões com o intuito de produzir dados novos e similares1. Dentro do âmbito empresarial e industrial, os modelos de IA generativa têm encontrado
aplicação em diversas áreas, como manufatura inteligente e sistemas de inteligência empresarial2. Diversas são as estratégias que utilizam a IA
para orientar os processos de aprendizagem de alta qualidade, causando grande impacto na educação no mundo3. O ensino superior foi
surpreendido pelo surgimento veloz e pelo crescimento exponencial na capacidade e uso da inteligência artificial generativa (IA). Talvez isso não
devesse ter sido ao caso, considerando a longa história de pesquisa no uso da IA em diversas áreas em técnicas de educação3. Podem
desempenhar um papel crucial no avanço dos recursos educacionais na área da educação médica, com potencial de complementar os métodos de
ensino convencionais e aprimorar a compreensão de conceitos médicos complexos4. Um