CBIS'24 XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Chatbots na identificação de problemas de amamentação: avaliação de desempenho

Chatbots in identification of breastfeeding issues: performance evaluation

Chatbots en la identificación de problemas de lactancia materna: evaluación del desempeño

Ari Pereira de Araújo Neto[1], Giovanny Rebouças Pinto[2], Joeckson dos Santos Corrêa[3], Liane Batista da Cruz Soares[4], Christyann Lima Campos Batista[5], Feliciana Santos Pinheiro[6], Ariel Soares Teles[7]

1 Mestre em Biotecnologia, Programa de Pós-Graduação em Biotecnologia, Universidade Federal do Delta do Parnaíba, Parnaíba (PI), Brasil.
2 Doutor em Ciências Biológicas, Programa de Pós-Graduação em Biotecnologia, Universidade Federal do Delta do Parnaíba, Parnaíba (PI), Brasil.
3 Mestre em Ciência da Computação, Programa de Pós-Graduação em Ciência da Computação, Universidade Federal do Maranhão, São Luís (MA), Brasil.
4 Mestra em Gestão de Programas e Serviços de Saúde, Banco de Leite Humano, Hospital Universitário da Universidade Federal do Maranhão, São Luís (MA), Brasil.
5 Doutor em Pediatria, Banco de Leite Humano, Hospital Universitário da Universidade Federal do Maranhão, São Luís (MA), Brasil.
6 Doutora em Pediatria, Departamento de Medicina III, Universidade Federal do Maranhão, São Luís (MA), Brasil.
7 Doutor em Engenharia Elétrica, Instituto Federal do Maranhão, Araioses (MA), Brasil.

Autor correspondente: MSc. Ari Pereira de Araújo Neto
E-mail: aripereiraneto@gmail.com
Link: (Arquivo Suplementar - https://doi.org/10.5281/zenodo.11377368)

Resumo

Objetivo: Este estudo objetivou avaliar o desempenho de *chatbots* de inteligência artificial na identificação de problemas relacionados à amamentação. Método: o estudo avaliou o *OpenAI ChatGPT3.5, Microsoft Copilot, Google Gemini e o Lhia* na identificação de problemas da amamentação. O chatbot *Lhia* está em desenvolvimento pelo nosso time de pesquisadores. Através do consenso entre profissionais de saúde especialistas em amamentação, foi criado um conjunto de dados de relatos de queixa clínica principal anotada em prontuários de atendimento do Hospital Universitário da Universidade Federal do Maranhão para os testes com três abordagens de comandos do tipo *zero-shot.* Resultados: o melhor desempenho foi com *ChatGPT-3.5*, que apresentou acurácia variando de 79% a 93%, *fallback* de 0% a 7% e *F1-score* de 75%

a 100%. Conclusão: *chatbots* de inteligência artificial podem ser uma ferramenta promissora para auxiliar mães e profissionais de saúde na detecção precoce de problemas na amamentação.

Descritores: Sistemas Inteligentes; Inteligência Artificial; Amamentação;

Abstract

Objective: This study aimed to evaluate the performance of artificial intelligence-based chatbots in identifying breastfeeding-related problems. Method: The study assessed *OpenAI ChatGPT-3.5*, *Microsoft Copilot, Google Gemini,* and *Lhia* in identifying breastfeeding issues. *Lhia* chatbot is being developed by our team of researchers. Through consensus among healthcare professionals specializing in breastfeeding, a dataset of annotated main clinical complaint reports from medical records at the University Hospital of the Federal University of Maranhão was created for testing with three zero-shot prompt approaches. Results: The best performance was achieved by *ChatGPT-3.5*, which demonstrated accuracy ranging from 79% to 93%, fallback from 0% to 7%, and F1-score from 75% to 100%. Conclusion: Artificial intelligence-based chatbots can be a promising tool to assist mothers and healthcare professionals in the early detection of breastfeeding issues.

Keywords: Expert Systems; Artificial Intelligence; Breastfeeding;

Resumen

Objetivo: Este estudio tuvo como objetivo evaluar el desempeño de chatbots de inteligencia artificial en la identificación de problemas relacionados con la lactancia. Metodo: El estudio evaluó *OpenAI ChatGPT-3.5*, *Microsoft Copilot, Google Gemini* y *Lhia* en la identificación de problemas de la lactancia. El chatbot *Lhia* está siendo desarrollado por nuestro equipo de investigadores. A través del consenso entre profesionales de salud especialistas en lactancia, se creó un conjunto de datos de informes de quejas clínicas principales anotadas en los registros médicos del Hospital Universitario de la Universidad Federal de Maranhão para las pruebas con tres enfoques de comandos del tipo *zero-shot*. Resultados: El mejor desempeño fue con *ChatGPT-3.5*, que presentó una precisión que varió del 79% al 93%, *fallback* del 0% al 7% y *F1-score* del 75% al 100%. Conclusión: Los chatbots de inteligencia artificial

pueden ser una herramienta prometedora para asistir a madres y profesionales de salud en la detección temprana de problemas en la lactancia.

Descriptores: Sistemas Especialistas; Inteligencia Artificial; Lactancia;

Introdução

O aleitamento materno exclusivo nos primeiros seis meses de vida é uma das intervenções mais eficazes para reduzir a mortalidade infantil. A Organização Mundial da Saúde (OMS) estima que a amamentação pode prevenir 823 mil mortes de crianças menores de cinco anos por ano.[1] Os benefícios da amamentação para a saúde humana vão desde a infância à vida adulta, sendo o leite materno o alimento ideal para o bebê, pois contém todos os nutrientes necessários ao seu crescimento e desenvolvimento. Além disso, a amamentação protege o bebê contra diversas doenças, tais como infecções respiratórias, diarreia, otites, alergias e obesidade.[2]

A amamentação, embora natural, apresenta desafios que podem levar ao desmame precoce, como dores durante a amamentação devido ao posicionamento e pega inadequados, fissuras mamilares, ingurgitamento, mastite, baixa produção de leite, entre outros problemas.[3] O desmame precoce representa riscos para a saúde da mãe e da criança, incluindo o aumento de infecções, alergias e obesidade.[2] Para prevenir o desmame precoce, é necessário o apoio da família, dos profissionais de saúde, bem como de políticas públicas para promoção do aleitamento materno.[4]

Os *chatbots* são soluções de software, também conhecidos como agentes conversacionais, pois se comunicam com o usuário simulando uma conversa de humano.[5] Os *chatbots* são classificados em dois tipos: baseados em regras e baseados em Inteligência Artificial (IA).[6] Os *chatbots* baseados em IA são capazes de entender a linguagem natural e não somente os comandos pré-definidos. Além disso, eles conseguem manter diferentes contextos de conversas e fornecer ao usuário conversas mais ricas e engajadas. [6]

Os *chatbots* têm desempenhado um papel cada vez mais significativo no apoio à amamentação, representando uma tecnologia promissora para permitir o acesso à informação e apoio às mães que amamentam.[7,8,9] A maioria das dúvidas relacionadas à amamentação estão enraizadas em mitos ou sistemas de crenças existentes nas redes de apoio das quais as mães fazem parte, e aproximadamente nove em cada dez dessas dúvidas podem ser respondidas por um *chatbot*.[9] As primíparas (recém-mães)

muitas vezes vivenciam ansiedade relacionada à amamentação e a intervenção de ferramentas como *chatbots* baseados em IA tem se mostrado ser uma excelente estratégia para oferecer informações sobre como lidar adequadamente com o problema que enfrentam.[8]

As interações com *chatbots* podem educar as mães sobre sua saúde, e seus médicos aprovam o uso deles como auxílio no monitoramento das mães durante os períodos pré e pós-natal.[8] Os *chatbots* baseados em IA já demonstraram potencial no apoio às questões relacionadas à educação em saúde,[10,11,12,13] assistência na resolução de problemas médicos,[14,15,16] inclusive às mães que enfrentam problemas de amamentação.[17]

Este estudo objetiva avaliar o desempenho dos *chatbots* baseados em IA *OpenAI ChatGPT-3.5, Microsoft Copilot, Google Gemini* e o chatbot *Lhia* na identificação de problemas relacionados à amamentação, com base na queixa clínica principal presente nos prontuários de atendimento especializado do Banco de Leite Humano do Hospital Universitário da Universidade Federal do Maranhão (BLH-HUUFMA). Este é um estudo pioneiro que envolve o processamento de linguagem natural por chatbots baseados em IA para fins de identificação de problemas de saúde materno-infantil no contexto do Português Brasileiro (PT-BR).

Métodos

Coleta de dados

A coleta de dados foi realizada em prontuários de mulheres atendidas no BLH-HUUFMA no período de janeiro de 2023 a janeiro de 2024. A rotina do atendimento especializado aos problemas da amamentação consistia na avaliação da saúde do binômio mãe-bebê, com avaliações da história clínica obstétrica, avaliação da amamentação sobre diferentes aspectos, anotando as principais queixas da mãe, hipóteses diagnósticas e condutas clínicas adotadas. Essas informações eram anotadas em prontuário do ambulatório do BLH-HUUFMA (Arquivo Suplementar 1 - https://doi.org/10.5281/zenodo.11377368). As consultas ocorreram de forma gratuita e sob livre demanda pelo Sistema Único de Saúde (SUS), sem agendamento prévio de acordo com o surgimento de problemas de amamentação. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa do HUUFMA (CAAE 63947022.0.0000.5086).

Para este estudo, foram selecionados os prontuários dos pacientes que continham, no campo 'queixa clínica principal', anotações dos profissionais que prestaram o atendimento clínico especializado e também apresentaram uma hipótese diagnóstica relacionada à queixa clínica do paciente. Após essa seleção, apenas os prontuários com hipóteses diagnósticas para os problemas de posição e pega, insegurança materna, ingurgitamento, fissura, mastite e hipogalactia foram avaliados, pois estes eram os problemas que o chatbot Lhia, em desenvolvimento desde 2022 no BLH-HUUFMA, era capaz de identificar.[17]

Anotação de dados

A partir desses registros, criamos um conjunto de dados contendo as sentenças relacionadas às queixas clínicas e suas respectivas hipóteses diagnósticas. Esse conjunto de dados resultante foi avaliado e validado por três profissionais de saúde (um médico pediatra, um fonoaudiólogo e um enfermeiro) com 9 a 42 anos de experiência clínica em atendimento especializado de problemas da amamentação, de modo a se chegar a uma concordância sobre aquela queixa clínica relatada e a respectiva hipótese diagnóstica do problema na amamentação. Além desse conjunto de queixas clínicas, foi feito um levantamento em conversas do *WhatsApp* em que as mães buscavam doar leite humano para o BLH-HUUFMA, representando uma classe no conjunto de dados que tratam da ausência de problemas na amamentação.

Testes com os *chatbots*

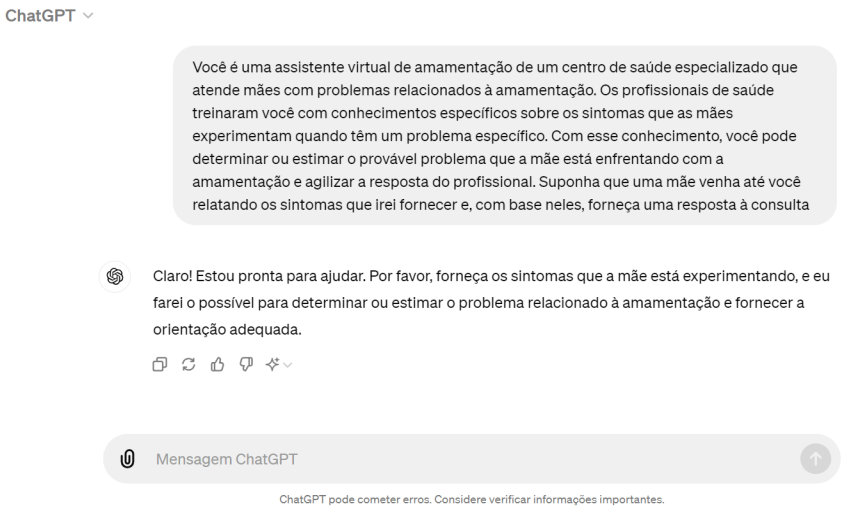
O conjunto de dados foi então usado como entrada para iniciar conversas e fornecer orientações com diferentes *chatbots* generativos: o *ChatGPT-3.5*, o *Copilot*, o *Gemini*, e o *Lhia*.[17] Este último é baseado em IA, possuindo um *Large Language Model* (LLM)[18], mas não é generativo. *Lhia* analisa a entrada do usuário para identificar a intenção da mensagem, com base em um modelo de classificação baseado em aprendizado de máquina, e retornar a melhor resposta correspondente de um conjunto predefinido de mensagens. Os demais *chatbots* (i.e,. *ChatGPT-3.5*, *Copilot*, e *Gemini*) usam LLMs generativos[18] com capacidade para gerar respostas a partir do texto de entrada.

Lhia é um *chatbot* baseado em modelo pré-treinado *DIET* (*Dual Intent and Entity Transformer*) atuando como classificador, incorporado ao modelo *BERT* (*Bidirectional*

Encoder Representations from Transformers), mais especificamente o *BERTimbau*. O *BERTimbau* é uma versão especializada do *BERT* para PT-BR, que foi treinada com dados da internet em português brasileiro. *Lhia* tem sido desenvolvido por nosso time de pesquisadores, sendo implementado utilizando a versão *BERTimbau large* com o *DIETClassifier* na plataforma *open source Rasa*[17]. Neste estudo, *Lhia* foi disponibilizado para interação via *WhatsApp*.

A capacidade dos *chatbots* generativos pode ser melhorada por meio da criação de comandos (do inglês, *prompts*) específicos, tornando os *chatbots* mais acessíveis e aplicáveis em diferentes domínios, aproveitando todo o potencial dos *chatbots.*[19] Neste estudo, adotamos a abordagem de comando *zero-shot*, ou seja, durante as perguntas, não foi fornecido nenhum exemplo de treinamento aos modelos. O uso da abordagem *zero-shot* é justificado porque buscamos fazer uma comparação justa entre o *Lhia* e os demais *chatbots*, uma vez que *Lhia* possui um modelo treinado apenas para a identificação de seis problemas relacionados à amamentação e a intenção de doar leite humano, envolvendo apenas queixas clínicas, sem outros parâmetros avaliados. Embora tenha sido adotada a técnica de comando *zero-shot*, a interação com os demais *chatbots* foi contextualizada pelo envio de uma mensagem no início de cada sessão de conversa (Figura 2). Sempre que a sessão terminava, este processo de contextualização era repetido.

Figura 2 – Mensagem inicial usada nos *chatbots* generativos.



Buscando otimizar o desempenho dos *chatbots* generativos no contexto da amamentação, em sessões diferentes, utilizou-se perguntas diferentes, denominadas aqui de comandos 01, 02 e 03 (Tabela 01).[13] Em diferentes sessões, para cada

sentença de queixa clínica, foi perguntado aos *chatbots* generativos a hipótese diagnóstica da seguinte maneira: introdução a sessão <relato da queixa clínica> comando 01; introdução a sessão <relato da queixa clínica> comando 02; e introdução a sessão <relato da queixa clínica> comando 03. Para o *chatbot Lhia* foi utilizado apenas o relato da queixa clínica, mas não foi utilizado qualquer comando, pois não era generativo e apenas identificava o problema exato de acordo com seu fluxo definido. Essa abordagem corresponde ao comando 01 para os *chatbots* generativos.

Tabela 1 – Estratégias de abordagens de comandos utilizados com os *chatbots* generativos.

Abordagem	Descrição do comando	Exemplo do comando
Comando 01 (problema exato)	Nessa abordagem era fornecida a queixa principal ao *chatbot* e questionado qual o exato problema	Existe relato de mamas cheias e endurecidas, qual o exato problema que a mãe está enfrentando?
Comando 02 (ranqueamento dos principais problemas)	Nessa abordagem era fornecida a queixa principal ao *chatbot* e solicitado o ranqueamento dos principais problemas	Existe relato de mamas cheias e endurecidas, ranqueia os principais problemas que a mãe possa estar enfrentando?
Comando 03 (opções de múltipla escolha)	Nessa abordagem era fornecida a queixa principal ao *chatbot* e dado as opções de escolha com os respectivos problemas	Existe relato de mamas cheias e endurecidas, entre os problemas abaixo, escolha aquele que a mãe está passando: a) posição e pega; b) fissura; c) ingurgitamento; d) mastite; e) insegurança materna; f) hipogalactia; g) nenhum dos anteriores.

As sentenças relacionadas às queixas clínicas foram organizadas em uma planilha (Arquivo Suplementar 2 - https://doi.org/10.5281/zenodo.11377368) com os seguintes dados: uma coluna referente à sentença, uma coluna para o problema rotulado pelos profissionais e colunas para registrar a hipótese de problema por cada *chatbot* em cada comando. Matrizes de confusão para cada *chatbot* e cada abordagem de comando em específico foram geradas para calcular as métricas de desempenho.[20] O desempenho dos modelos foi analisado de acordo com as seguintes métricas: acurácia, precisão, *recall* e *F1-score*.[20] O índice de *fallback* foi calculado a partir da relação dos números de gatilhos de *fallback* com o número total de interações. Os gatilhos de *fallback* ocorreram quando o *chatbot* não entendeu a entrada ou deu uma resposta alternativa, não identificando o problema.[17]

Resultados

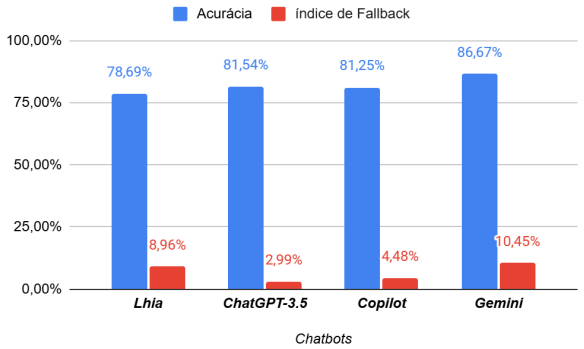
Conjunto de dados

No período avaliado, 159 binômios mãe-filho foram atendidos pelo serviço especializado do ambulatório de amamentação do BLH-HUUFMA por apresentarem problemas relacionados à amamentação e tiveram suas informações de atendimento anotadas nas fichas. Após a reunião de consenso para a seleção das fichas que apresentavam os relatos mais comuns para cada problema específico, um conjunto de dados final foi composto por 57 sentenças relatando queixas clínicas relacionadas aos problemas e 10 sentenças coletadas de conversas do aplicativo *WhatsApp*, relatando a intenção de doar leite humano. O conjunto de dados apresentava 10 frases para o problema de posição e pega, 10 para ingurgitamento, 10 para fissura, 10 para mastite, 09 para hipogalactia, 08 para insegurança materna e 10 para a intenção de doar leite.

Desempenho dos *chatbots*

A Figura 3 apresenta os resultados de acurácia dos *chatbots* utilizando a abordagem de comando 01. O *Gemini* obteve a maior acurácia ao identificar corretamente 86,67% das queixas clínicas. O *ChatGPT-3.5* identificou corretamente 81,54% das queixas, seguido por *Copilot* (81,25%) e *Lhia* (78,69%). Todos forneceram gatilhos de *fallback*, mas o pior desempenho nessa métrica foi o *Gemini*, com um índice de *fallback* de 10,45% das interações, seguido do *Lhia* com 8,96%, *Copilot* 4,48%, e *ChatGPT-3.5* com 2,99%.

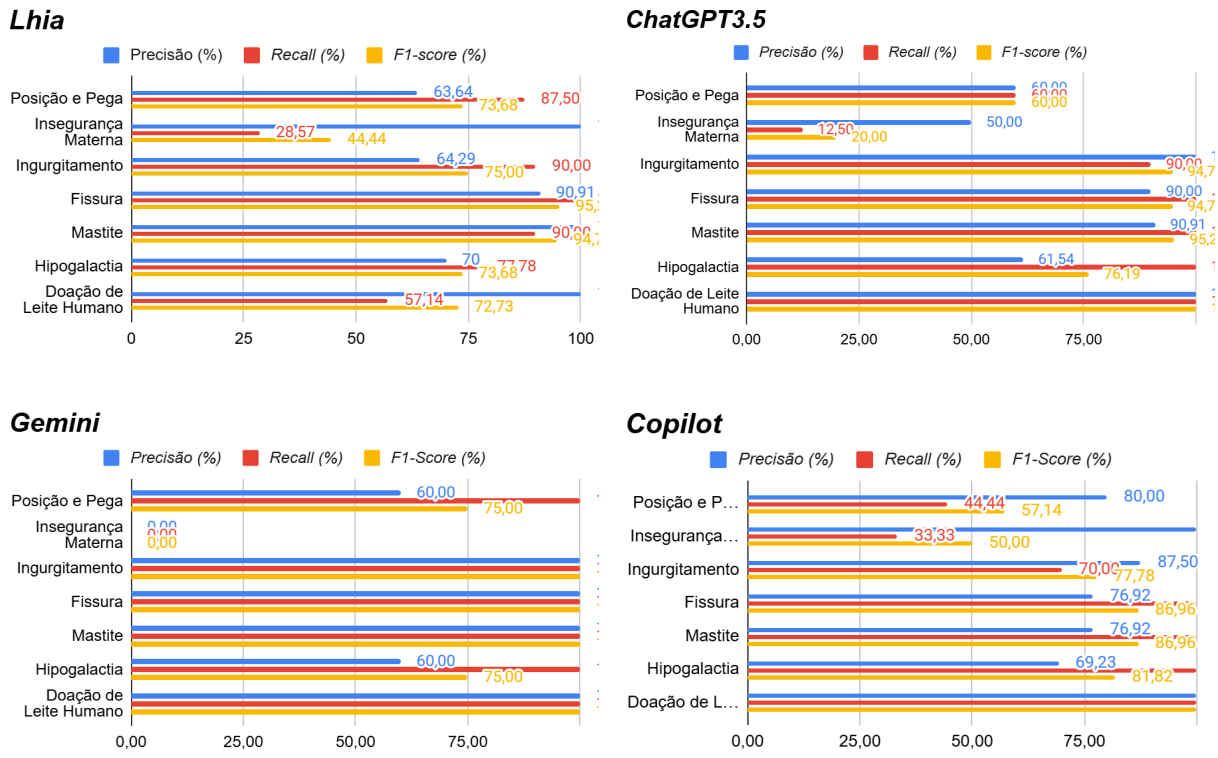
Figura 3 – Acurácia e índice de *fallback* dos *chatbots* na abordagem de comando 01.



A Figura 4 mostra os resultados de desempenho dos quatro *chatbots* em relação à precisão, *recall* e *F1-score* com a interação na abordagem de comando 01. O *Gemini* obteve o melhor desempenho na maioria dos problemas, porém, foi incapaz

de identificar as queixas clínicas referentes à insegurança materna. O *Copilot* superou todos os outros *chatbots* na detecção de insegurança materna, com F1-score de 50%. O *ChatGPT-3.5* também obteve excelente desempenho na detecção da maioria dos problemas (*F1-score* maior que 94,74%), com desempenho mediano para hipogalactia e posição e pega (F1-score de 60%) e desempenho ruim na detecção de insegurança materna (F1-score de 20%). O *Lhia* apresentou resultados excelentes para a detecção de fissura e mastite (*F1-score* de 94,74%), desempenho mediano para hipogalactia, ingurgitamento e posição e pega (*F1-score* de 73,68% a 75%), e desempenho ruim para detecção de insegurança materna (*F1-score* de 44,44%). Na identificação da intenção de doar leite humano, todos os *chatbots* apresentaram desempenho excelente (F1-score de 100%), exceto o *Lhia* que apresentou desempenho mediano (F1-score de 72,73%).

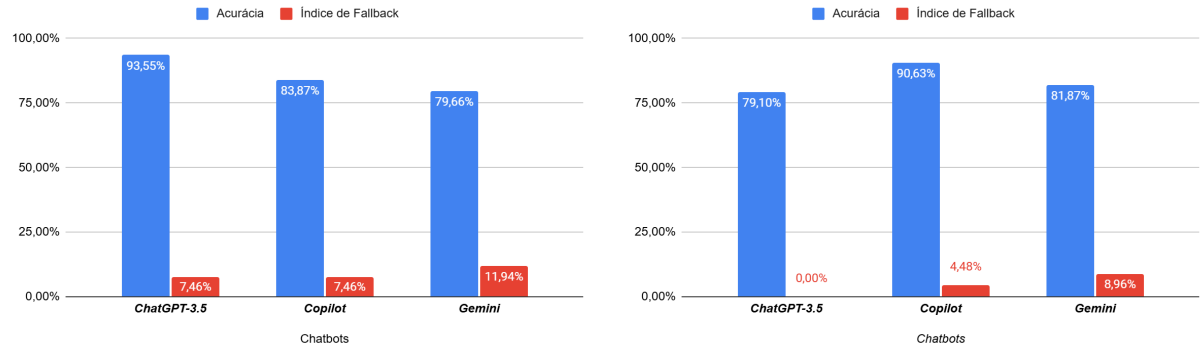
Figura 4 – Precisão, *recall* e F1-score dos chatbots na abordagem de comando 01.



A Figura 5 mostra o desempenho dos *chatbots* generativos para as abordagens de comando 02 e 03, nas quais não se avaliou o *chatbot Lhia.* Na abordagem de comando 02, o *ChatGPT-3.5* apresentou melhor desempenho, com 93,55% de acurácia, seguido do *Copilot* com 83,87% e *Gemini* com 79,66%. Na abordagem de

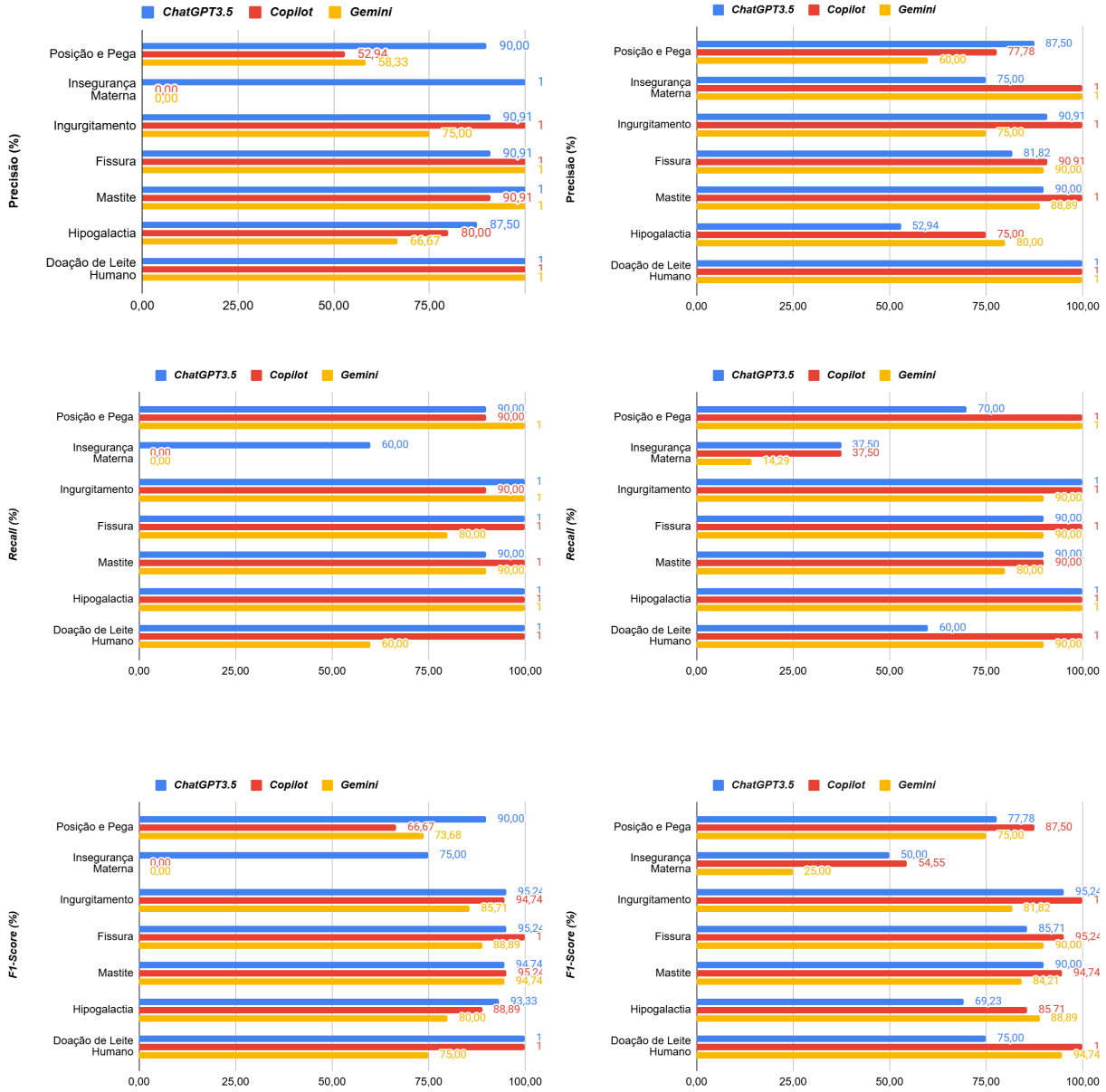
comando 03, o *Copilot* apresentou melhor desempenho, com 90,63% de acurácia, seguido do *Gemini* com 81,97% e *ChatGPT-3.5* com 79,10%.

Figura 5 – Acurácia e índice de *fallback* dos *chatbots* nas abordagens de comando 02 e 03.



A Figura 6 mostra os resultados de desempenho dos *chatbots* generativos em relação à precisão, *recall* e *F1-score* com as interações nas abordagens de comando *02 e 03.* No comando 02, no qual o *ChatGPT-3.5* obteve excelente desempenho na identificação da maioria dos problemas (*F1-score* de 90% a 100%), com desempenho mediano na classificação das queixas clínicas referentes à insegurança materna (*F1-score* de 75%). O *Copilot* apresentou desempenho semelhante (*F1-score* de 88,89 a 100%), mas com desempenho mediano na identificação de problemas na posição e pega (*F1-score* de 66,67%), e não foi capaz de identificar queixas relacionadas à insegurança materna. Da mesma forma, o *Gemini* foi incapaz de identificar insegurança materna, porém teve desempenho excelente para a detecção de fissura e mastite (*F1-score* de 88,89% e 94,74%) e desempenho mediano para hipogalactia, ingurgitamento e posição e pega (*F1-score* de 73,68 a 85,71%). Na identificação de quando a mãe não apresentava problema, tanto o *ChatGPT-3.5* quanto o *Copilot* apresentaram desempenho excelente (*F1-score* de 100%), no entanto, o *Gemini* apresentou desempenho mediano (*F1-score* de 75%).

Figura 6 – Precisão, *recall* e *F1-score* dos chatbots nas abordagens de comando 02 e 03.



Discussão

Na avaliação de desempenho dos *chatbots* realizada neste estudo, o *ChatGPT-3.5* apresentou o desempenho geral mais alto, ao ter os melhores resultados para acurácia e índice de *fallback*. Por outro lado, o *Lhia* teve o menor desempenho. Dentre os *chatbots* avaliados neste estudo, apenas o *Lhia* foi treinado com dados relacionados aos problemas específicos da amamentação durante seu desenvolvimento em 2022. No entanto, o modelo utilizado no *Lhia*, baseado no BERTimbau e DIET, não é generativo e possui uma quantidade muito inferior de

parâmetros, quando comparado aos *chatbots* generativos avaliados. Dessa forma, o seu desempenho mais baixo pode estar relacionado a sua habilidade de generalização.

O *ChatGPT-3.5* atingiu acurácia de 93,55% e *F1-score* entre 90 e 100% no comando 02, o qual solicitava a indicação ranqueada de prováveis problemas, uma situação mais próxima da sua utilização clínica. Na prática clínica, os problemas de amamentação estão geralmente inter-relacionados e há mais de um problema existente. Portanto, não há usualmente o diagnóstico exato de apenas um problema, como foi solicitado nas abordagens de comando 01 e 03. Por esse motivo, os resultados encontrados para a abordagem de comando 02 devem ser considerados mais relevantes, quando analisando os *chatbots* generativos.

O *ChatGPT-3.5* tem apresentado resultados promissores na utilização em diferentes áreas da saúde. Na avaliação das respostas de *chatbots* às perguntas sobre audiologia feitas por estudantes e médicos especialistas, o *ChatGPT* alcançou a pontuação geral mais alta, enquanto o *Bard*, versão anterior do *chatbot* da Google, obteve a mais baixa.[21] Na oftalmologia, resultados semelhantes foram encontrados ressaltando o potencial dos *chatbots*, particularmente do *ChatGPT-4.0*, para fornecer informações precisas e respostas abrangentes às perguntas relacionadas à miopia.[22] Na neurologia, a colaboração entre profissionais de saúde e os *chatbots* têm o potencial de melhorar o diagnóstico diferencial, com melhores resultados quando os profissionais utilizam o auxílio do *ChatGPT-4.0* para a análise das ressonâncias magnéticas cerebrais.[23]

Na abordagem de comando *zero-shot,* tem sido demonstrado desempenho superior do *ChatGPT-4.0* em comparação com *ChatGPT-3.5* e *Bard*.[19] Em nosso estudo, resultados semelhantes foram obtidos na abordagem de comando 03, quando foram dadas opções de múltipla escolha referentes aos problemas. Nesta mesma abordagem, o *Copilot*, o qual possui internamente uma versão do LLM *OpenAI GPT-4.0*, obteve melhor desempenho, exceto para a detecção de insegurança materna. Neste estudo, excetuando o *ChatGPT-3.5*, todos os outros *chatbots* não apresentaram desempenho satisfatório para a detecção de insegurança materna, mesmo para as diferentes abordagens de comandos. Acreditamos que isso ocorreu, possivelmente, por se tratar de um problema complexo de ser identificado e por

envolver diferentes variáveis de avaliação na clínica médica.[24] Apesar disso, o *ChatGPT-3.5* apresentou o melhor desempenho na classificação das queixas clínicas referentes à insegurança materna.

Esse estudo possui limitações que devem ser reconhecidas. Primeiro, não foram abordados a grande diversidade de problemas que podem acontecer na amamentação, mas sim aqueles mais frequentemente encontrados na realidade do BLH-HUUFMA, podendo não ser a realidade de diferentes populações. Segundo, a seleção das queixas principais gerou dados desbalanceados, com menor quantidade de exemplos para insegurança materna e hipogalactia, o que pode ter influenciado no percentual de desempenho dos chatbots na identificação desses problemas. Outra limitação do estudo é que, buscando uma comparação justa com o *chatbot Lhia*, utilizamos apenas a abordagem de comando *zero-shot* para os *chatbots* generativos, podendo ser abordadas outras formas de comando, tais como *few-shot* e cadeia de pensamento, as quais fornecem mais informações prévias ao *chatbot*.[18]

Conclusão

Esse estudo objetivou avaliar o desempenho de *chatbots* baseados em IA na identificação de problemas relacionados à amamentação, com base na queixa clínica principal presente nos prontuários de atendimento especializado do BLH-HUUFMA. Concluímos haver potencial da tecnologia *chatbot* baseado em IA na identificação de problemas na amamentação. No entanto, é crucial reconhecer que o uso de *chatbots* na saúde materno-infantil ainda enfrenta desafios significativos, tais como a complexidade e variedade das situações clínicas que estão além daquelas testadas neste estudo. Apesar disso, este estudo contribui para a base de conhecimento emergente sobre o uso de *chatbots* na saúde materno-infantil e destaca oportunidades para desenvolvimento dessas tecnologias, visando melhorar a qualidade do suporte oferecido por essas ferramentas inovadoras em saúde. Como trabalho futuro, integraremos o *chatbot Lhia* ao *ChatGPT* e avaliaremos os aspectos de usabilidade e experiência do usuário, envolvendo mães que já amamentam (puerpério) e que ainda não amamentam (pré-natal). Além disso, planejamos avaliar o *chatbot* como uma ferramenta de educação em saúde tanto com mães quanto com profissionais de saúde não especializados.

Agradecimentos

Agradecimentos à Unidade de Obstetrícia do HU-UFMA e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (308059/2022-0).

Referências

1. Victora CG, Bahl R, Barros AJ, França GV, Horton S, Krasevec J, et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. The Lancet. 2016;387(10017):475-90.

2. Rollins NC, Bhandari N, Hajeebhoy N, Horton S, Lutter CK, Martines JC, et al. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. Lancet. 2016;387(10017):491-504

3. World Health Organization. Infant and Young Child Feeding: Model Chapter for Textbooks for Medical Students and Allied Health Professionals; Technical Report; WHO: Geneva, Switzerland, 2009

4. Bhattacharjee NV, Schaeffer LE, Hay SI, Lu D, Schipp MF, Lazzar-Atwood A, et al. Mapping inequalities in exclusive breastfeeding in low- and middle-income countries 2000–2018. Nat Hum Behav. 2021; 5,1027–1045.

5. Softić A, Husić JB, Softić A, Baraković S. Health chatbot: design, implementation, acceptance and usage motivation. In 20th International Symposium Infoteh-Jahorina; 2021 17-19 March; East Sarajevo, Bosnia and Herzegovina. IEEE; 2021, pp. 1-6, Available from: doi:10.1109/INFOTEH51037.2021.9400693.

6. Prasad VA, Ranjith R. Intelligent chatbot for lab security and automation. In 11th international conference on computing, communication and networking Technologies, 2020 1-3 July; Kharagpur, India. IEEE, 2020, pp. 1-4, Available from: doi:10.1109/ICCCNT49239.2020.9225641.

7. Yadav D, Malik P, Dabas K, Singh P. Feedpal: Understanding Opportunities for Chatbots in Breastfeeding Education of Women in India. Proceedings of the ACM on Human-Computer Interaction. 2019;3(CSCW):170:1–30.

8. Gupta V, Arora N, Jain Y, Mokashi S, Panda C. Assessment on Adoption Behavior of First-time Mothers on the Usage of Chatbots for Breastfeeding Consultation. J Mahatma Gandhi Univ Med Sci Tech. 2021;6(2):64-68.

9. Montenegro JL, Costa CA, Janssen LP. Evaluating the use of chatbot during pregnancy: A usability study. Healthcare Analytics. 2022;2(100072):1-9.

10. Campos-Filho AS, Cursino JR, Barros-Júnior TD, Lima EC. Assistente Virtual na Educação em Saúde dos Homens. J Health Inform. 2023;15(Esp):1-14.

11. Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatry. 2023;22(3):479-480.

12. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative ai assistance: qualitative study of popular large language models. JMIR Medical Education. 2024;*10*(1), 1-11.

13. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use chatgpt for mental health and substance use education? examining its quality and potential harms. JMIR Medical Education. 2023;*9*(1), 1-10.

14. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv. 2023:2(2303.13375), 1-35.

15. Lautrup AD, Hyrup T, Schneider-Kamp A, Dahl M, Lindholt JS, Schneider-Kamp P. Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. Open Heart. 2023;10(2), 1-8.

16. Andrew A. Potential applications and implications of large language models in primary care. Fam Med Community Health, 2024;12(Suppl 1), 1-6.

17. Corrêa JS, Araújo-Neto AP, Pinto GR, Lima LD, Teles AS. Lhia: a smart chatbot for breastfeeding education and recruitment of human milk donors. Appl Sci. 2023;13(12): 1-19

18. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv preprint arXiv: 2023:13(2303.18223), 1-124.

19. Espejel JL, Ettifouri EH, Alassan MS, Chouham EM, Dahhane W. GPT-3.5, GPT-4, or BARD? evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. Natural Language Processing Journal. 2023;5(100032), 1-192.

20. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv preprint arXiv: 2020:1(2008.05756), 1-17.

21. Jedrzejczak WW, Kochanek K. Comparison of the audiological knowledge of three chatbots-ChatGPT, Bing Chat, and Bard. Audiol Neurootol. 2023:11(38710158), 1-7.

22. Lim ZW, Pushpanathan K, Yew SM, Lai Y, Sun CH, Lam JS, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023:95(104770), 1-11.

23. Kim SH, Schramm S, Berberich C, Rosenkranz E, Schmitzer L, Serguen K, et al. Human-AI collaboration in large language model-assisted brain mri differential diagnosis: a usability study. medRxiv 2024:02(05), 1-22.

24. Simas WL, Penha JS, Soares LB, Rabêlo PP, Oliveira BL, Pinheiro FS. Insegurança materna na amamentação em lactantes atendidas em um banco de leite humano. Rev. Bras. Saude Mater. Infant. 2021:21(1), 251-259.