MarIA - DeepSeek: Uma Proposta de Assistente por Modelo Amplo de Linguagem para Agentes Comunitários de Saúde

Pedro A. F. França[1], Rafaela V. P. Sá[3], Silas Alves-Costa[3], Poliana C. de A. F. Viola[3], Sérgio S. Costa[3], Bruno F. de Souza,[3] João D. S. de Almeida[1], João O. B. Diniz[12], Cecilia C. C. Ribeiro[3]

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
Av. dos Portugueses, 1966 - Vila Bacanga, São Luís - MA, 65080-805

[2]Fábrica de Inovação – Instituto Federal do Maranhão (IFMA)
BR-226, S/N – Vila Nova - Grajaú – MA, 65940-000

[3]Pós-Graduação em Saúde Coletiva – Universidade Federal do Maranhão (UFMA)
R. Barão de Itapari, 155 - Centro, São Luís - MA, 65020-070

{pedro.franca@nca.ufma.br}

Abstract. *The first 1,000 days of life, including pregnancy and the child's first two years, represent a critical window for preventing non-communicable chronic diseases (NCDs). However, Community Health Agents (CHAs) face challenges in accessing and applying evidence-based recommendations during this period. This work presents MarIA - DeepSeek, a virtual assistant powered by large language models (LLMs), which integrates techniques such as Prompt Chaining, Retrieval-Augmented Generation (RAG), and expert-curated documents to deliver personalized, evidence-based guidance to CHAs. Experiments showed that the tool improves the accuracy, clarity, and accessibility of recommendations, outperforming general-purpose models like GPT-4.0 and Gemini, and enhancing decision-making in primary health care.*

Resumo. *Os primeiros 1.000 dias de vida, que compreendem a gestação e os dois primeiros anos da criança, representam um período crítico para prevenir doenças crônicas não transmissíveis (DCNT). No entanto, Agentes Comunitários de Saúde (ACS) enfrentam dificuldades para acessar e aplicar recomendações baseadas em evidências científicas durante esse período. Este trabalho apresenta o MarIA - DeepSeek, uma assistente virtual baseada em modelos amplos de linguagem (LLMs), que integra técnicas de Prompt Chaining, Retrieval-Augmented Generation (RAG) e curadoria de documentos especializados para oferecer orientações personalizadas e cientificamente embasadas aos ACS. Os experimentos demonstraram que a ferramenta aprimora a precisão, a clareza e a acessibilidade das recomendações, superando modelos generalistas como GPT-4.0 e Gemini, e contribuindo para uma tomada de decisão mais eficaz na atenção primária à saúde.*

1. Introdução

Os primeiros 1.000 dias de vida englobam a gestação e os dois primeiros anos da criança, são o período de maior plasticidade do fenótipo humano, com intensa atividade neurológica, imunológica e metabólica, influenciando fortemente o desenvolvimento infantil [DOHaD-SAP et al. 2020]. O conceito DOHaD (do inlês Developmental Origins of

Health and Disease), Origens Desenvolvimentistas da Saúde e das Doenças, expandiu a compreensão dos primeiros 1000 dias como o período crítico do desenvolvimento, mostrando que fatores de risco sociais, ambientais e nutricionais agindo nesta fase tem repercussões a longo prazo, aumentando o risco de doenças crônicas não transmissíveis (DCNT) no futuro [DOHaD-SAP et al. 2020, Alves-Costa et al. 2024].

Durante a gestação, iniquidades sociais, estresse, tabagismo, excesso de peso, consumo de álcool e de dieta não saudável elevam o risco de complicações na gravidez, como hipertensão e diabetes gestacional [Alves-Costa et al. 2024]. Em sequência, somam-se outros fatores de risco, como parto cesáreo, nascimento pré-termo, tempo de amamentação insuficiente, introdução precoce de açúcares e uso de antibióticos, resultando no aparecimento das primeiras DCNT na infância: obesidade, cárie, asma e alergias [Araújo et al. 2024, Muniz et al. 2022, Nascimento et al. 2017]. Estas doenças não estão apenas associadas entre si, mas também aumentam o risco de outras DCNT mais mortais no futuro [Majbauddin et al. 2019].

Assim, os primeiros 1000 dias são uma janela de oportunidades para promover intervenções mais efetivas para prevenção das DCNT, com redução substancial de risco as ações voltadas para a vida intrauterina [Alves-Costa et al. 2024]. Os Agentes Comunitários de Saúde (ACS) desempenham um papel essencial na promoção da saúde da gestante, especialmente por meio de visitas domiciliares, agendamento das consultas de pré-natal, identificação das gestantes em vulnerabilidade, garantindo encaminhamentos adequados e continuidade do cuidado [Bonifácio et al. 2019].

Entretanto, a qualificação dos ACS enfrenta desafios: embora o conteúdo abranja de forma ampla a área da saúde, o processo formativo geralmente se baseia em cursos introdutórios e treinamentos informais [Bonifácio et al. 2019]. Soma-se o acesso limitado a informações baseadas em evidências científicas por estes profissionais, o que pode comprometer as ações preventivas mais efetivas nos primeiro 1000 dias de vida.

Neste contexto, integrar tecnologias aos cuidados na saúde pode promover a atenção básica mais resolutiva, capaz de identificar, classificar riscos, e oportunizar intervenções mais eficazes na atenção integral a saúde materna e infantil (Política Nacional de Atenção Integral à Saúde da Criança). A Inteligência Artificial (IA) surge como uma ferramenta promissora que vem sendo utilizada na área da saúde [Diniz et al. 2021, Diniz et al. 2024] , e que pode fortalecer a atuação dos ACS nas recomendações baseadas em evidências científicas para prevenção das primeiras DCNT nos 1000 dias de vida.

Este trabalho propõe o desenvolvimento da ferramenta MarIA - DeepSeek, um assistente virtual baseado em IA projetado para apoiar os ACS nas recomendações baseadas em evidências científicas para prevenção das primeiras DCNT. Espera-se que essa solução contribua significativamente para aprimorar a tomada de decisão de ACS por um assistente baseado em IA, facilitando o acesso a informações atualizadas e baseadas em evidências científicas. Este estudo traz a utilização inédita do DeepSeek como LLM (Large Language Model) destinado à atenção primária à saúde, promovendo uma abordagem mais personalizada.

2. Trabalhos Relacionados

Os métodos computacionais e ferramentas baseadas em LLM já são amplamente pesquisados para suportar tarefas na área da saúde. Esta seção descreve alguns assistentes

virtuais recentes encontrados na literatura.

Em [Passinato et al. 2024], propôs-se um chatbot oftalmológico baseado em modelos de código aberto e técnicas de Geração Aumentada de Recuperação (*Retrieval-augmented generation* - RAG), sem ajuste fino, para facilitar o acesso a informações sobre saúde ocular. Foram testadas três abordagens de RAG, utilizando o Mistral 7B como modelo gerador e o e5-multilingual como indexador. A avaliação, conduzida com o framework Ragas e o ChatGPT como crítico, analisou a relevância do contexto e da resposta. Os resultados mostraram que todas as técnicas superaram o GPT-3.5 em relevância da resposta. Isso demonstra que técnicas RAG com modelos *open-source* são viáveis e eficazes, oferecendo uma alternativa acessível para chatbots especializados em saúde.

O estudo de [Rodrigues et al. 2024] desenvolveu um chatbot para a Atenção Primária à Saúde, implementado na plataforma *ManyChat* e integrado ao Telegram, em uma Unidade de Saúde da Família em Pernambuco. O chatbot fornecia informações sobre serviços de saúde, prevenção de doenças e um canal para sugestões dos pacientes. Os resultados indicaram uma boa aceitação, evidenciando o potencial das tecnologias de informação para aprimorar o suporte na APS.

[Cardenas et al. 2024] introduz o *AutoHealth*, um sistema de *Internet of Medical Things* (IoMT) para o gerenciamento personalizado da Doença de Parkinson, utilizando IA. Integrando *smartwatches*, aplicativos móveis e um chatbot baseado em LLMs, o sistema monitora continuamente sintomas como tremores e congelamento da marcha. O *AutoHealth* coleta e processa dados de movimento e voz com aprendizado de máquina, oferecendo *feedback* personalizado em tempo real. O chatbot auxilia pacientes com orientações sobre medicamentos, exercícios e bem-estar emocional.

Assim como as soluções mencionadas, a MarIA surge da necessidade de integrar LLM e métodos avançados de recuperação de informação para aprimorar a experiência dos usuários em diversos contextos de saúde. Diferentemente de outros chatbots, o MarIA - DeepSeek inova ao contextualizar cada resposta com dados científicos e recomendações personalizadas, utilizando uma linguagem adaptada aos ACS. Dessa forma, contribui para ampliar a cobertura dos serviços de atenção básica e potencializa a detecção precoce de riscos, promovendo o acesso a protocolos atualizados e fortalecendo o Sistema Único de Saúde (SUS) no cenário materno-infantil brasileiro.

3. Materiais e Método

O método proposto (Figura 1), inicia-se com a **Predição da Calculadora de Risco**, que estima a probabilidade de complicações gestacionais. Em seguida, a **Seleção de Fatores da Gestante** filtra variáveis relevantes (histórico clínico, sinais vitais, exames etc.), recebidas de uma calculadora externa. Este trabalho foca exclusivamente no modelo de Processamento de Linguagem Natural (PLN) que é alimentada por um modulo de calculadora que envia as informações de probabilidade de desfecho e fatores de risco associados.

Os dados alimentam o **Prompt Inicial**, que estabelece o contexto para o modelo de linguagem. A técnica de *Prompt Chaining* permite refinar respostas e gerar novas consultas. Paralelamente, a **Recuperação** busca evidências clínicas e protocolos atualizados via DeepSeek [Guo et al. 2025]. Os documentos utilizados passam por curadoria de especialistas em saúde, garantindo materiais científicos de qualidade. A ferramenta também é escalonável, permitindo a inclusão de novos documentos com evidências.

Por fim, o sistema gera uma **Recomendação** com orientações clínicas, formuladas de modo acessível para a ACS comunicar-se claramente com a gestante. Caso necessário, uma **Nova Pergunta** reinicia o ciclo, aprimorando continuamente as recomendações.

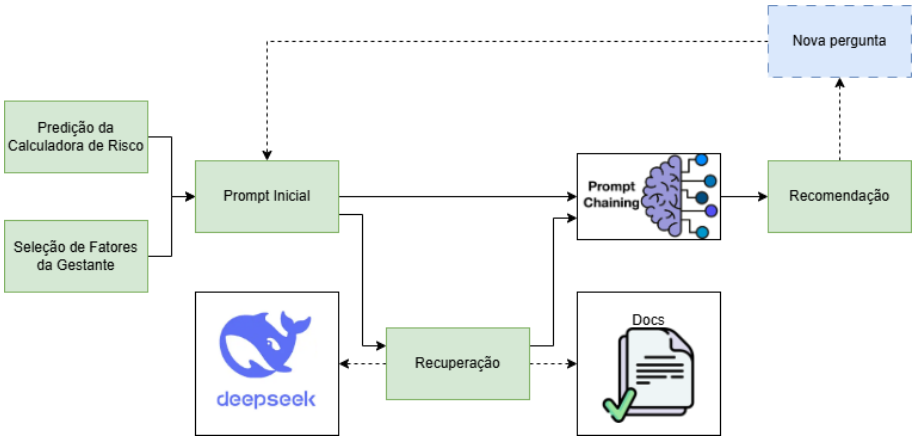

Figura 1. Método Proposto

4. Conjunto de Dados

O conjunto de dados foi criteriosamente construído a partir de artigos científicos, periódicos especializados e estudos de coorte que abordam os quatro desfechos-alvo da calculadora de risco (Obesidade, Alergia, Asma e Cárie). Esses materiais fundamentam a etapa de Seleção de Fatores da Gestante e são utilizados no Prompt para contextualizar a recomendação personalizada. Todo o conteúdo foi analisado por uma equipe composta por 3 odontologistas e 2 nutricionistas, garantindo a validade técnica. A abordagem quali-quantitativa permitiu tanto mensurar evidências como interpretar padrões subjetivos de linguagem nas respostas geradas.

Além disso, a integração de bases externas ampliou a cobertura semântica e a robustez do modelo, permitindo respostas contextualizadas com evidências científicas verificadas [Gao et al. 2023].

4.1. Entrada de Dados

A entrada de dados é crucial para sistemas interativos, permitindo interações dinâmicas e aprendizado contínuo [Cursino et al. 2020]. Na MarIA, seu funcionamento como chatbot depende de *inputs* que direcionam as respostas e interações com os ACS. A principal fonte de dados é a calculadora de risco, que utiliza análises preditivas para identificar fatores associados a doenças crônicas, como asma, alergia, cárie e obesidade. Esses dados são oriundos da Coorte Pré-Natal BRISA, conforme descrito por [da Silva et al. 2014].

A MarIA não realiza previsões diretamente, apenas consome os dados gerados pela calculadora, que é desenvolvida com base em nos fatores de riscos associados a cada desfecho [Nascimento et al. 2017]. Cada *input* inicial, proveniente da integração dos fatores identificados, possibilita não apenas a avaliação de riscos, mas também fundamenta as interações subsequentes entre a agente inteligente e os ACS. Para garantir a segurança e integridade dos dados, todas as informações são transmitidas de forma segura pelo *backend*, assegurando que as respostas geradas reflitam fielmente os dados processados e a evolução contínua do sistema.

4.2. Representação e Geração de Texto

4.2.1. Embeddings e Representação Semântica

As representações vetoriais das palavras são fundamentais no PLN, capturando características semânticas e sintáticas para melhorar o desempenho de modelos [Souza et al. 2020]. Modelos baseados em *Transformers*, como os LLMs, utilizam técnicas avançadas, como *subword* e *contextual embeddings*, que refinam a representação linguística e ampliam a capacidade de generalização dos modelos [Souza et al. 2020].

A MarIA adota o modelo *BERTimbau* (NeuralMind) [Souza et al. 2020], treinado especificamente para o português, garantindo melhor preservação semântica e adaptação à língua. Testes indicaram que modelos em inglês apresentaram instabilidades e *underfitting*, reforçando a escolha de uma solução otimizada para o português.

O BERTimbau possui uma arquitetura bidirecional, considerando o contexto anterior e posterior de uma palavra para gerar representações mais ricas e precisas. Como resultado, os *embeddings* extraídos das últimas camadas do modelo são altamente adaptáveis a diferentes contextos, fortalecendo sua aplicação em PLN [Devlin 2018].

4.2.2. Vetorização da Base de Dados

A utilização de bancos de dados vetoriais permite buscas eficientes por similaridade, sendo essencial para sistemas de LLM e IA generativa. Essa abordagem estrutura dados vetoriais para buscas semânticas, comparando os *embeddings* das consultas com documentos armazenados por meio de métricas de similaridade, garantindo a recuperação do conteúdo mais relevante [Sakai et al. 2024].

Na MarIA, os *embeddings* gerados a partir de dados sobre saúde materno-infantil são armazenados no ChromaDB, permitindo buscas rápidas e precisas. O ChromaDB ranqueia os resultados por similaridade e recupera dados pré-processados, fornecendo contexto para aprimorar as respostas dos modelos de linguagem. Esse fluxo otimiza a análise das perguntas dos ACS, garantindo acesso rápido a informações científicas sobre fatores de risco, prevenção e cuidados nos primeiros 1000 dias de vida, fortalecendo sua atuação no acompanhamento das gestantes e suas famílias.

4.2.3. Modelo de Linguagem e Geração de Respostas

Após a recuperação dos documentos via busca semântica, é necessário interpretar e estruturar as informações antes de apresentá-las ao usuário. Para isso, a MarIA utiliza o modelo DeepSeek-R1, que combina aprendizado por reforço e *Mixture of Experts* para aprimorar a tomada de decisão e adaptar-se a domínios específicos [Mikhail et al. 2025].

A geração de respostas é estruturada pela *Conversational Retrieval Chain*, que integra a recuperação de informações à geração adaptativa de respostas. O histórico de interações é armazenado na *ConversationBufferMemory*, configurada no módulo *langchain.memory* para referenciar diálogos anteriores e evitar repetições. Isso permite a formulação de respostas mais coesas e contextuais [Roy et al. 2025].

O DeepSeek organiza as informações extraídas e adapta a resposta ao perfil do usuário, garantindo explicações relevantes. Essa abordagem melhora a experiência do usuário, proporcionando diálogos mais naturais e alinhados às necessidades dos ACS.

4.2.4. Retrieval-Augmented Generation (RAG)

O *Retrieval-Augmented Generation* (RAG) aprimora LLMs ao integrar um mecanismo de recuperação de informações, mitigando as limitações de modelos puramente paramétricos. Esse processo permite que o modelo acesse fontes externas antes da geração da resposta, aumentando sua precisão e contextualização [Roy et al. 2025].

O RAG opera em três etapas: indexação, recuperação e geração. Na indexação, os documentos são segmentados em *chunks* e transformados em vetores de *embeddings* por meio de modelos como o *BERT*, armazenados em dados vetoriais, como o *ChromaDB*. Na recuperação, a consulta é convertida em um vetor de *embeddings*, e os fragmentos mais relevantes são identificados pelo cálculo da similaridade, garantindo a seleção de informações adequadas. Enquanto na geração, os fragmentos recuperados são incorporados ao *prompt*, que combina essas informações com seu conhecimento paramétrico para formular respostas coerentes e contextualizadas. Esse método, conhecido como *Retrieve-Read*, reduz alucinações, melhora a confiabilidade das respostas [Gao et al. 2023].

Na MarIA, o RAG é otimizado para o português, utilizando o *BERTimbau* [Souza et al. 2020]. A recuperação adota uma similaridade de 0.8 para selecionar apenas os *chunks* mais relevantes, que são então processados pelo modelo *DeepSeek* para produzir respostas mais precisas e contextualizadas. Para garantir transparência, incorporam-se metadados, como fonte e número de página, vinculando cada resposta às suas origens.

4.3. Engenharia de Prompt

Esta etapa é essencial para otimizar a interação com LLMs, garantindo respostas mais precisas, coerentes e alinhadas aos objetivos desejados. Além disso, contribui para mitigar vieses, imprecisões e a dependência excessiva da IA [Heston and Khun 2023].

Na MarIA, essa técnica adapta informações técnicas para um formato acessível aos ACSs, utilizando uma persona cuidadosamente desenvolvida para estabelecer um tom adequado. A linguagem é ajustada ao nível de escolaridade dos usuários, tornando o conhecimento científico mais compreensível e aplicável ao cotidiano das gestantes e comunidades atendidas. Algumas técnicas de Engenharia de *prompt* foram utilizadas para melhorar nossos resultados e são descritas a seguir.

4.3.1. Prompt Chaining

O *Prompt Chaining* é uma técnica que divide tarefas em uma sequência de *prompts* interligados, aprimorando a coerência e precisão das respostas geradas por LLMs [Wei et al. 2022]. Na MarIA, essa abordagem inicia com um *prompt* baseado em um modelo fixo, no qual variáveis como desfecho e fatores de risco são ajustadas dinamicamente. Esse formato garante padronização e flexibilidade para diferentes cenários clínicos.

A resposta é estruturada progressivamente em três etapas. Primeiro, a persona MarIA elabora uma explicação acessível sobre a relação entre o desfecho e seus fatores de risco baseado em RAG. Em seguida, um segundo *prompt* converte essa explicação em diretrizes práticas para os ACSs. Por fim, um terceiro *prompt* mantém a interação contínua, respondendo a dúvidas com base no contexto estabelecido. Essa estrutura organiza a comunicação, garantindo maior clareza, consistência e aplicabilidade das respostas.

4.3.2. Adaptação Dinâmica e Novas Interações

A MarIA aprimora a interação dos ACSs ao viabilizar um diálogo contínuo e contextualizado. Após gerar uma resposta detalhada ao *prompt* inicial, o sistema permite que novas perguntas sejam formuladas sem reinicializar o contexto, graças à Memória Conversacional, que preserva e ajusta dinamicamente o histórico das interações.

Além disso, a recuperação de informações em um banco vetorial complementa as respostas com dados validados, garantindo maior precisão. A integração entre *prompt chaining*, Memória Conversacional e busca vetorizada permite interações mais naturais e adaptáveis às necessidades dos ACSs e das gestantes, tornando o suporte técnico mais eficiente e alinhado à atenção primária à saúde.

5. Resultados e Discussão

Esta seção apresenta os resultados experimentais. Inicialmente, descreve-se o ambiente de desenvolvimento, seguido dos experimentos quali-quantitativos para validar o modelo. Por fim, o *framework* é analisado, abordando seus impactos e limitações.

5.1. Ambiente de desenvolvimento

O método foi desenvolvido em linguagem *Python*. Utilizou-se principalmente as bibliotecas *PyTorch* e *Transformers* da *HuggingFace* para o processamento de linguagem natural, integradas ao *framework Langchain* para a construção e gerenciamento de cadeias conversacionais. A API (*Application Programming Interface*) do *DeepSeek* foi empregada para a geração de respostas contextuais. O computador utilizado para os experimentos consiste num dispositivo equipado com um processador AMD Ryzen 5 2600X de 3.60ĠHz, 24,0ĠB de RAM, rodando em um sistema operacional Windows 11 Pro.

5.2. Experimentos

Nesta seção, serão apresentados testes de validação para mensurar as respostas.

5.2.1. Teste A/B para Avaliação das Respostas

Para validar a qualidade das respostas pelo modelo MarIA, realizou-se um teste A/B comparando seu desempenho com os modelos ChatGPT (OpenAI) e Gemini (Google). O objetivo foi verificar a capacidade de cada modelo em fornecer recomendações precisas para ACS, considerando os desfechos analisados neste estudo. A avaliação considerou oito recomendações, distribuídas entre nos quatro desfechos clínicos (obesidade, asma, alergia e cárie), combinados com seus fatores de risco. As respostas geradas foram analisadas

por cinco observadores independentes da área da saúde, com base nos seguintes critérios: clareza e compreensibilidade, precisão das informações e Relevância da recomendação.

Os resultados foram organizados com base na frequência de escolhas, permitindo identificar a abordagem mais alinhada às necessidades dos ACSs. Seguindo princípios de experimentação controlada [Kohavi et al. 2009], os testes no GPT-4.0 e no Gemini foram realizados em contas anônimas para evitar vieses. A Tabela 1 apresenta os achados.

Tabela 1. Resultados do Teste A/B. IA1 representa a MarIA, IA2 o GPT-4.0 e IA3 o Gemini.

Desfecho	Fatores	Maior Frequência	Percentual de Escolha
Asma	Ocupação mão de obra não qualificada, Obesidade pré-gestacional, Nascimento pré-termo, Parto cesárea, Periodontites na gestação	IA1	80%
Asma	Hipertensão na gestação, Consumo de álcool na gestação, Consumo semanal de ultraprocessados, Anemia na gestação, Consumo diário de refrigerante	IA1	60%
Alergia	Ocupação mão de obra não qualificada, Obesidade pré-gestacional, Nascimento pré-termo, Parto cesárea, Periodontites na gestação	IA1	60%
Alergia	Hipertensão na gestação, Consumo de álcool na gestação, Consumo semanal de ultraprocessados, Anemia na gestação, Consumo diário de refrigerante	IA1	100%
Cárie	Ocupação mão de obra não qualificada, Consumo diário de refrigerante na gestação, Consumo semanal de ultraprocessados, Periodontites na gestação, Obesidade pré-gestacional	IA1	100%
Cárie	Consumo de álcool na gestação, Hipertensão na gestação, Parto cesárea, Nascimento pré-termo, Anemia na gestação	IA1	80%
Obesidade	Ocupação mão de obra não qualificada, Obesidade pré-gestacional, Nascimento pré-termo, Parto cesárea, Periodontites na gestação	IA1	100%
Obesidade	Hipertensão na gestação, Consumo de álcool na gestação, Consumo semanal de ultraprocessados, Anemia na gestação, Consumo diário de refrigerante	IA1	100%

A análise mostra que respostas da MarIA, com RAG para o domínio da saúde, contribuiu para sua maior aceitação. Nos desfechos de cárie e obesidade, a escolha do MarIA foi unânime, indicando a adaptação do modelo ao contexto dos ACS. Por outro lado, nos desfechos de asma e alergia, onde há maior volume de evidências científicas, modelos generalistas como GPT-4.0 e Gemini também foram selecionados como melhor resposta em algumas avaliações. Além disso, fatores como hipertensão gestacional, consumo de álcool e alimentação ultraprocessada foram recorrentes entre os determinantes dos desfechos analisados, e a capacidade do MarIA de adaptar suas recomendações a essas variáveis reforça seu potencial como ferramenta de suporte à tomada de decisão.

5.2.2. Cosseno de similaridade

O cosseno de similaridade é uma métrica utilizada para medir a proximidade entre vetores, sendo útil na comparação de textos gerados por modelos LLM [Manning and Schütze 2008]. Seus valores variam de 0 a 1, onde valores mais altos indi-

cam maior similaridade. Aplicou-se essa métrica para comparar as respostas dos modelos MarIA, GPT-4.0 e Gemini. A Tabela 2 apresenta os resultados.

Tabela 2. Similaridade entre as respostas dos modelos.

Comparação de Modelos	Cosseno de Similaridade
MarIA vs. GPT-4.0	0.7699
MarIA vs. Gemini	0.7952
GPT-4.0 vs. Gemini	0.9041

Observa-se que o GPT-4.0 e o Gemini apresentam a maior similaridade (0,90), sugerindo que suas respostas são altamente similares. MarIA exibe menor similaridade com ambos (0,76), confirmando que suas respostas são distintas. Esse achado corrobora os resultados do teste A/B, onde os observadores tenderam a escolher a MarIA, indicando que suas respostas são de fato diferentes das geradas por GPT-4.0 e Gemini. Isso reforça a importância de modelos personalizados, treinados com bases científicas.

5.3. Framework

Nesta seção, é apresentado o funcionamento da MarIA. A Figura 2 descreve suas etapas.

Figura 2. Interface do assistente MarIA em diferentes estados.

Na Figura 2 (A), observa-se o status inicial, no qual a recomendação está sendo gerada a partir dos dados recebidos da calculadora. Esse momento representa o processamento das informações antes da exibição dos resultados.

Na Figura 2 (B), o assistente exibe explicações detalhadas sobre os fatores de risco identificados, permitindo que o ACS compreenda melhor as condições associadas aos desfechos e suas explicações. Essa etapa é essencial para fundamentar a tomada de decisão e possibilitar um acompanhamento mais preciso da gestante.

Por fim, na Figura 2 (C), são apresentadas dicas práticas destinadas a auxiliar os ACS na comunicação com a gestante. Essas sugestões visam facilitar a orientação sobre os cuidados necessários, reforçando a adoção de hábitos saudáveis e a prevenção de problemas relacionados aos fatores de risco identificados.

Além disso, ao final da interface, há uma caixa de texto que permite novas interações, possibilitando que o usuário faça perguntas adicionais e refine as informações recebidas. De maneira geral, a interface do MarIA é simples e intuitiva, proporcionando uma experiência fluida e acessível para os ACS durante o atendimento.

5.3.1. Impactos e Limitações

O assistente possui uma interface simples e intuitiva, permitindo que o ACS obtenha recomendações imediatas. Além de descrever os fatores de risco, o modelo oferece orientações claras para facilitar a comunicação com a gestante, traduzindo informações médicas complexas em recomendações acessíveis e baseadas em evidências, o que melhora o aconselhamento e apoia a tomada de decisões na Atenção Primária à Saúde. Entre os impactos esperados pelo projeto, destaca-se o aprimoramento do suporte ao ACS, proporcionando recomendações mais embasadas. A recuperação de dados em banco vetorial facilita o acesso a protocolos e diretrizes atualizadas, promovendo maior acessibilidade às informações de saúde. Além disso, a personalização do atendimento é viabilizada pelo armazenamento do histórico de respostas, tornando as recomendações mais contextuais e individualizadas, enquanto o fortalecimento técnico do ACS é favorecido ao garantir acesso contínuo a orientações atualizadas durante as visitas domiciliares.

Apesar dos benefícios, algumas limitações devem ser consideradas. A dependência de infraestrutura tecnológica pode representar um desafio, especialmente em regiões com acesso limitado à internet e recursos computacionais. Além disso, há o risco de imprecisão ou viés nas respostas devido a lacunas nos dados de treinamento ou limitações inerentes a sistemas baseados em LLM.

6. Conclusão

O assistente MarIA demonstrou-se uma ferramenta promissora, possibilitando recomendações personalizadas no atendimento materno-infantil. A utilização de LLM, aliada a técnicas como *Prompt Chaining* e RAG, permitiu um aprimoramento na contextualização das respostas, superando as limitações de abordagens generalistas.

A interface foi projetada para ser intuitiva e acessível, facilitando a interação com a ferramenta. No entanto, desafios como a necessidade de avaliação contínua da precisão das recomendações e a adaptação do modelo às especificidades regionais persistem.

Como trabalhos futuros, pretende-se ampliar a base de conhecimento com novas diretrizes clínicas e integrar uma interface para interações por voz. Além disso, após validada com profissionais da saúde, serão conduzidos estudos para avaliar o impacto na prática dos ACS e sua aceitação na atenção primária pelas gestantes.

Agradecimentos

Agradecemos o Ministério da Saúde/DECIT/CNPq 400759/2024-1 junto com Gates Fundation. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil, Code-001, PGPG-Amazônia Legal (88887.637402/2021-00) e PROCAD Amazônia (88881.719704/2022-01). Bolsa de Produtividade em Pesquisa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

Alves-Costa, S., da S. Pereira, S. M., Haddad, A. E., Ribeiro, C. C. C., and Oliveira, A. E. F. (2024). *The First Thousand Days of Life: A DOHaD Perspective to Dentistry*. EDUFMA, 1st edition.

Araújo, S. M. P., Nascimento, G. G., Ladeira, L. L. C., Alves-Costa, S., Saraiva, M. C., Alves, C. M. C., Thomaz, E. B. A. F., and Ribeiro, C. C. C. (2024). Chronic oral disease burden at the first 1000 days: Intergenerational risk factors, brisa cohort. *Oral Diseases*, 30(8):5388–5396.

Bonifácio, L. P., Marques, J. M., and Vieira, E. M. (2019). Assessment of the knowledge of brazilian community health workers regarding prenatal care. *Primary Health Care Research & Development*, 20:e21.

Cardenas, L., Parajes, K., Zhu, M., and Zhai, S. (2024). Autohealth: Advanced llm-empowered wearable personalized medical butler for parkinson's disease management. In *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0375–0379. IEEE.

Cursino, J. R. V., Calista, A. A., Nascimento, J. E. d. M., and Campos Filho, A. S. d. (2020). Uma revisão integrativa sobre o uso de chatbot para subsidiar o ensino na área da saúde. *Revista de Saúde Digital e Tecnologias Educacionais*.

da Silva, A. A. M., Simões, V. M. F., Barbieri, M. A., Cardoso, V. C., Alves, C. M. C., Thomaz, E. B. A. F., de Sousa Queiroz, R. C., Cavalli, R. C., Batista, R. F. L., and Bettiol, H. (2014). A protocol to identify non-classical risk factors for preterm births: the brazilian ribeirão preto and são luís prenatal cohort (brisa). *Reproductive Health*, 11:1–9.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diniz, J. O., Dias Jr, D. A., da Cruz, L. B., Marques, R. C., Gomes Jr, D. L., Cortês, O. A., de Carvalho Filho, A. O., and Quintanilha, D. B. (2024). Efficientensemble: Diagnóstico de câncer de mama em imagens de ultrassom utilizando processamento de imagens e ensemble de efficientnets. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 202–213. SBC.

Diniz, J. O., Ferreira, J. L., da Silva, G. L., Quintanilha, D. B., Silva, A. C., and Paiva, A. (2021). Segmentação de coração em tomografias computadorizadas utilizando atlas probabilístico e redes neurais convolucionais. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 83–94. SBC.

DOHaD-SAP, S., de la Salud, O., et al. (2020). Developmental origins of health and disease concept: The environment in the first 1000 days of life and its association with noncommunicable diseases. *Archivos argentinos de pediatria*, 118(4):S118–S129.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Heston, T. F. and Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, 2(3):198–205.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18:140–181.

Majbauddin, A., Tanimura, C., Aoto, H., Otani, S., Parrenas, M. C., Kobayashi, N., Morita, T., Inoue, K., Masumoto, T., and Kurozawa, Y. (2019). Association between dental caries indicators and serum glycated hemoglobin-levels among patients with type 2 diabetes mellitus. *Journal of oral science*, 61(2):335–342.

Manning, C. D. and Schütze, H. (2008). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Mikhail, D., Farah, A., Milad, J., Nassrallah, W., Mihalache, A., Milad, D., Antaki, F., Balas, M., Popovic, M. M., Feo, A., et al. (2025). Performance of deepseek-r1 in ophthalmology: An evaluation of clinical decision-making and cost-effectiveness. *medRxiv*, pages 2025–02.

Muniz, A. K. O. A., Ribeiro, C. C. C., Vianna, E. O., Serra, H. C. O. A., Nascimento, J. X. P. T., Cardoso, V. C., Barbieri, M. A., Da Silva, A. A. M., and Bettiol, H. (2022). Factors associated with allergy traits around the 2nd year of life: a brazilian cohort study. *BMC pediatrics*, 22(1):703.

Nascimento, J. X. P. T., Ribeiro, C. C. C., Batista, R. F. L., de Britto Alves, M. T. S. S., Simões, V. M. F., Padilha, L. L., Cardoso, V. C., Vianna, E. O., Bettiol, H., Barbieri, M. A., et al. (2017). The first 1000 days of life factors associated with "childhood asthma symptoms": Brisa cohort, brazil. *Scientific reports*, 7(1):16028.

Passinato, E. B., Rios, W. S., and Galvão Filho, A. R. (2024). Integração de modelos de linguagem e rag na criação de chatbots oftalmológicos. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 354–365. SBC.

Rodrigues, R. R., dos Santos Vasconcellos, A., and de Campos Filho, A. S. (2024). Desenvolvimento de um chatbot na atenção primária à saúde. *Journal of Health Informatics*, 16(Especial).

Roy, S., Goswami, M., Nargund, N., Mohanty, S., and Pattnaik, P. K. (2025). Conversational text extraction with large language models using retrieval-augmented systems. *arXiv preprint arXiv:2501.09801*.

Sakai, K., Uehara, Y., and Kashihara, S. (2024). Implementation and evaluation of llm-based conversational systems on a low-cost device. In *2024 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 392–399. IEEE.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.