Integração de Modelos de Linguagem e RAG na Criação de Chatbots Oftalmológicos

Emanuel B. Passinato[1], **Walcy S. R. Rios**[1], **Arlindo R. Galvão Filho**[1]

[1]Centro de Competência Embrapii de Tecnologias Imersivas – AKCIT
Universidade Federal de Goiás
Goiânia – GO – Brasil

{emanuel.passinato , walcy.rios}@discente.ufg.br, arlindo@ufg.br

Abstract. *Accessibility to ophthalmological services is an important factor in determining eye health, being influenced by the socioeconomic status of individuals. To facilitate access to information about eye health, recent works in the field focus on using established private language models or those with fine-tuning, both approaches involving additional costs, whether financial, data base needs, or complexity. This study proposes the development of a chatbot using open-source language models and retrieval augmented generation (RAG) techniques. Three techniques were evaluated naive RAG, HYDE and Rewrite-Retrieve-Read. To evaluate the retrieved context and the generated response, ChatGPT was used as a critic through the Ragas framework. The results indicate that it is possible to surpass the baseline performance of GPT-3.5 with the proposed techniques, reducing costs and attesting to the viability of similar projects.*

Resumo. *A acessibilidade aos serviços oftalmológicos é um fator importante para determinar a saúde ocular, sendo influenciada pelo estado socioeconômico dos indivíduos. Para facilitar o acesso às informações sobre saúde ocular, trabalhos recentes na área focam em utilizar modelos de língua já consolidados de mercado ou com ajuste fino, ambas as abordagens apresentam custos extras, seja financeiro, necessidade de base de dados ou complexidade. Este estudo propõe o desenvolvimento de um chatbot utilizando modelos de linguagem de código aberto e técnicas de retrieval augmented generation (RAG), sem ajuste fino. Três técnicas foram avaliadas, naive RAG, HYDE e Rewrite-Retrieve-Read. A avaliação do sistema RAG, foi realizada utilizando o ChatGPT como modelo crítico, por meio do framework Ragas. Os resultados indicam que é possível superar a performance base do GPT-3.5 com as técnicas propostas, reduzindo custos e atestando a viabilidade de projetos similares.*

1. Introdução

A oftalmologia desempenha um papel vital na manutenção e no aprimoramento da saúde ocular, afetando diretamente a qualidade de vida dos indivíduos [Assi et al. 2021]. O olho humano, uma estrutura complexa e delicada, requer cuidados especializados para prevenir e tratar uma vasta gama de condições que podem comprometer a visão. A importância dessa especialidade médica vai além do tratamento de doenças, abrangendo a prevenção de problemas visuais e a promoção da saúde ocular. Diante dos avanços tecnológicos

e científicos, a oftalmologia tem experimentado progressos significativos, ampliando as possibilidades de diagnóstico precoce e tratamentos mais eficazes, o que sublinha a necessidade de acesso universal a tais serviços para garantir o bem-estar e a inclusão social dos indivíduos.

No entanto, a acessibilidade aos serviços oftalmológicos permanece desigual, sendo profundamente influenciada pelo estado socioeconômico. Estudos têm demonstrado consistentemente que a posição socioeconômica está diretamente relacionada à procura por cuidados oftalmológicos, com indivíduos em situações de maior vulnerabilidade econômica apresentando menor tendência para buscar esses serviços essenciais. A baixa conscientização em saúde ocular, associada a uma menor adesão às diretrizes para exames oftalmológicos, destaca a necessidade crítica de estratégias inclusivas que abordem as disparidades no acesso aos cuidados de saúde ocular [Organization 2019].

Com a popularização das inteligências artificiais generativas, em particular dos modelos de língua como o *ChatGPT*, a sociedade se viu em meio a uma nova era tecnológica. O sucesso do modelo da *OpenAI* ficou evidente logo após seu lançamento, alcançando a marca de 100 milhões de usuários ativos em 2 meses.

Porém, modelos de língua ainda possuem limitações claras, em especial, quando utilizados para tarefas que exigem conhecimento intensivo [Lewis et al. 2020], como é o caso de temas relacionados à medicina. Em tarefas de conhecimento intensivo, os modelos atuais tendem a alucinar, gerando respostas incorretas e prejudicando os usuários. Outro problema latente, é o sigilo dos dados no uso de serviços externos, visto que os dados da área da saúde tem naturalmente sigilo.

As recentes pesquisas na área [Antaki et al. 2023], [Bernstein et al. 2023], exploram a utilização de modelos de mercado como o *ChatGPT* , ou modelos com ajuste fino como em [Zhao et al. 2023]. As soluções que envolvem a utilização de modelos pagos trazem uma carga econômica considerável para o produto final, podendo impactar na sua distribuição e utilização, especialmente para pessoas de baixa renda. Por outro lado, modelos com ajuste fino necessitam de um conjunto considerável de dados para treinar e avaliar o modelo, dificultando a rápida implementação e adicionando custos de pesquisa.

Neste cenário, propomos a utilização de modelos de código aberto, juntamente com a técnica *Retrieval Augmented Generation* (RAG) [Lewis et al. 2020], visando obter um modelo que seja factual, acessível e mais eficiente do que modelos de mercado ao responder perguntas sobre problemas oftalmológicos, sem a necessidade de ajuste fino. Três técnicas de RAG foram avaliadas, a versão original proposta pelo trabalho [Lewis et al. 2020], doravante referenciada como *naive* RAG; *Hypothetical Document Embeddings* (HYDE) [Gao et al. 2023] e *Rewrite-Retrieve-Read* [Ma et al. 2023].

A metodologia foi dividida em quatro etapas: 1) Aquisição da base de conhecimento sobre oftalmologia; 2) Utilização de técnicas de RAG para fornecer suporte ao modelo gerador; 3) Geração e avaliação das respostas e do contexto recuperado; e 4) Análise de sensibilidade ao tamanho do contexto em quatro etapas: 250, 500, 750 e 1000 caracteres.

O objetivo do presente trabalho é realizar uma análise detalhada e abrangente sobre a viabilidade da aplicação de técnicas avançadas de *Retrieval Augmented Generation* (RAG) na criação de *chatbots* especializados na área da saúde. Esta pesquisa visa não ape-

nas examinar a eficácia dessas técnicas, mas também explorar suas possíveis limitações e identificar oportunidades para aprimoramento.

2. Trabalhos Relacionados

Em [Bernstein et al. 2023] realizou um estudo comparativo entre respostas geradas pelo *ChatGPT* e profissionais oftalmologistas em perguntas sobre cuidado ocular. A forma avaliativa do estudo trata-se de uma comparação em 200 pares de perguntas e respostas classificadas em duas categorias: gerada por máquina ou por humano. A acurácia na categorização entre as respostas foi de 61.3%, onde o autor constatou que a máquina preferia textos muito longos, facilitando assim a categorização. É destacado ainda que as respostas geradas pelo *ChatGPT* não se diferenciam consideravelmente dos especialistas no quesito da veracidade da informação ou desvios dos padrões da comunidade oftalmologista.

Em [Antaki et al. 2023] avaliou a capacidade das primeiras versões do modelo de língua *ChatGPT*, em cenário de múltipla escolha em dois conjuntos de dados comuns sobre oftalmologia, *Ophthalmic Knowledge Assessment Program (OKAP)* e *Basic and Clinical Science Course* (BCSC). A principal métrica foi a acurácia pelo fator de múltipla escolha. Os resultados reportados foram de 59.4% e 49.2% respectivamente. O autor constata uma pontuação satisfatória no teste OKAP, com pontuação considerada alta. Para possível melhoria é recomendado a continuação da fase de pré-treino para o domínio oftalmológico.

Em [Zhao et al. 2023] focou na especialização da arquitetura *Llama-2* de 7 bilhões de parâmetros para o cenário de dados oftalmológicos a partir de um conjunto de dados privado de aproximadamente 70 mil amostras. A principal métrica utilizada para avaliação foi a métrica ROUGE, feita com cálculos de sobreposição de n-gramas entre o rótulo esperado. O autor constatou melhora significativa em relação aos outros modelos de língua genéricos. Além disso, é discutido as limitações da métrica selecionada para a avaliação, pois baseia-se primariamente na correspondência direta entre o texto gerado e o rótulo esperado. Contudo, diferentes profissionais podem oferecer diagnósticos diferentes e ponderamentos particulares.

3. Materiais e Métodos

3.1. Conjunto de dados

Para o desenvolvimento de um sistema RAG é essencial que este possua um conjunto de documentos de referência que serão utilizados como contexto para o modelo gerador. O presente estudo foi conduzido utilizando uma base de dados extraídos manualmente da internet, de artigos estilo blog escritos e mantidos por hospitais e clínicas especializadas em tratamentos oftalmológicos. Ao todo 714 documentos foram utilizados como fonte de referência.

Ao conduzirmos a avaliação sobre a performance do sistema, utilizamos um conjunto de 40 perguntas, simulando perguntas de pacientes. As perguntas utilizadas nos experimentos foram redigidas com base nos próprios documentos recuperados, selecionadas e revisadas pelos autores, visando abranger uma variedade de temas, diferentes estilos de texto e dificuldade de responder.

3.2. Estratégias de avaliação

Avaliação qualitativa em geração de texto aplicado não é trivial [Gao et al. 2024a], pois diversos fatores a influenciam, como: subjetividade, fluidez, relevância, coerência, entre outros. Métricas de avaliação tradicionais como *BLEU* [Papineni et al. 2002] e *ROUGE* [Lin 2004] baseiam-se na sobreposição de *ngrams* entre texto gerado e a referências para medir sua qualidade. Alguns resultados recentes mostram que essas métricas possuem baixa correlação com julgamentos humanos [Sulem et al. 2018], portanto não podem avaliar o texto de forma confiável. Ao incluir técnicas de recuperação de texto para aprimorar a geração (RAG), é essencial que também o texto recuperado seja avaliado, especialmente sobre o quão relevante ele é para a assertividade da resposta. [Es et al. 2024] propôs técnicas de automatização de métricas qualitativas através da interação com outro modelo gerador mais capacitado sendo análogo à função de crítico nas respostas geradas.

Neste trabalho serão utilizado três principais métricas para avaliação da solução, são elas: 1) fidelidade ao contexto (em inglês: *faithfulness*), com o propósito de mensurar o quanto a resposta gerada foi pautada no contexto fornecido; 2) relevância da resposta, com foco em mensurar o quão pertinente a resposta é, com base na pergunta; 3) relevância do contexto, com foco de avaliar o quão pertinente o contexto recuperado é, para responder a pergunta. A implementação da avaliação teve como base o *framework Ragas* [Es et al. 2024]

3.3. *Retrieval Augmented Generation* - RAG

Neste estudo foram exploradas três abordagens de RAG: *naive* RAG, *hypothetical document embeddings* (HYDE) e *Rewrite-Retrieve-Read* (RRR). O foco das técnicas citadas é na etapa de seleção/recuperação de documentos. A seleção de documentos dá-se através do cálculo da similaridade do cosseno entre os vetores latentes da pergunta e dos documentos fornecidos como base de dados. Os modelos indexadores recebem um texto como entrada e fornecem como saída um vetor que caracteriza aquele texto, criando uma estrutura de dados chave-valor, com a chave sendo um elemento vetorial e o valor sendo o texto original.

3.3.1. *Naive RAG*

Os documentos mais relevantes para responder à entrada do usuário são selecionados por similaridade. Utilizando-se da própria entrada (*query*) como ponto de partida, o modelo indexador gera um vetor para essa entrada e compara com as chaves (*keys*) da base vetorial. Após a recuperação dos documentos mais similares, estes são compostos juntamente com a entrada inicial do usuário para a formação de um novo *prompt*. Este *prompt* é por fim enviado ao modelo de língua para a geração. Esse processo visa fornecer ao modelo gerador, as informações necessárias para responder à pergunta do usuário, reduzindo alucinações e aumentando a relevância da resposta.

3.3.2. *Hypothetical Document Embeddings (HYDE)*

HYDE [Gao et al. 2023] é uma técnica avançada de RAG, que busca resolver um problema comum à abordagem do *naive RAG*, este problema é: assimetria entre a pergunta

e o documento de contexto para respondê-la. Essa assimetria possui várias fontes, sendo as principais: diferença de tamanho das sentenças, diferença de estilo de texto e ruídos da pergunta. A consequência dessa assimetria é uma imprecisão na recuperação dos melhores documentos possíveis.

Para resolver esses problemas, a proposta do HYDE é aproximar o texto de entrada do *retriever* com os documentos previamente indexados, gerando um documento hipotético a partir da entrada do usuário. O documento hipotético gerado pode ainda conter alucinações, porém será um texto mais similar ao documento esperado do que utilizar puramente a entrada do usuário. De posse do texto hipotético o *retriever* recupera o documento, ou conjunto de documentos, mais similar para compor o *prompt* final de geração da resposta.

3.3.3. *Rewrite-Retrieve-Read* (RRR)

A ideia central do *Rewrite-Retrieve-Read* é parecida com a do HYDE, reescrever o texto que será utilizado pelo *retriever* para a busca dos documentos. A diferença se encontra na forma do texto reescrito. No *framework Rewrite-Retrieve-Read* é extraída da pergunta original uma ou mais perguntas independentes para a busca dos conhecimentos necessários na base vetorial. Esse processo visa retirar possíveis ruídos da pergunta, além de separar temas quando a pergunta abrange mais de um tema, para uma busca independente na base de conhecimento.

3.4. Modelo gerador: Mistral

O modelo escolhido para gerar as respostas foi o Mistral [Jiang et al. 2023], um modelo de linguagem de 7 bilhões de parâmetros, projetado visando desempenho e eficiência. O modelo utiliza *grouped-query attention* (GQA) para inferência mais rápida, juntamente com *sliding window attention* (SWA) para lidar com sequências de entrada de comprimento variável, reduzindo o custo de inferência. Esse modelo provou ser uma excelente alternativa, performando melhor que modelos maiores como o modelo 13B (Llama 2) e em alguns casos melhor que o modelo 34B (Llama 1).

3.5. Modelo indexador: e5-multilingual

E5 [Wang et al. 2024] é um modelo indexador que pode ser utilizado como um modelo de *embedding* de uso geral, para quaisquer tarefas que exijam uma representação vetorial única de textos, como: recuperação (*retrieval*), categorização e classificação. O E5 alcançou forte desempenho tanto em *zero shot* quanto em *fine tuning*. Em seu lançamento o E5 obteve os melhores resultados no *benchmark* MTEB, superando os modelos indexadores existentes com 40x mais parâmetros.

3.6. Experimentos

Para a condução dos experimentos, primeiramente foi construída uma linha de base, utilizando um modelo de mercado, o *GPT-3.5-turbo* (*ChatGPT*). A linha de base foi construída utilizando a API da OpenAI, sem informar ao modelo qualquer contexto, apenas a pergunta foi utilizada para a geração. Portanto, apenas métricas relacionadas à resposta podem ser aplicadas para a linha de base.

Pergunta	Resposta
Posso tomar banho de rio ou de piscina com lente de contato?	Não recomendo usar lentes de contato durante banhos em rios ou piscinas, pois há um risco significativo de bactérias e outros microorganismos que podem causar infecções no olho. Para garantir a higiene ocular, é melhor remover as lentes antes de entrar em água e usá-las apenas quando estiverem limpas e descascadas. Em vez disso, você pode considerar usar óculos protectores para nadar ou proteção solar adequada se estiver no sol

Tabela 1. Exemplo de resposta correta

Pergunta	Resposta
Não usar óculos, faz o grau aumentar??	Sim, se você não usar óculos quando é necessário, o grau de correção pode aumentar sobre o tempo. Isso ocorre porque as lentes de óculos corrigem deficiências visuais específicas e, se essas deficiências continuam sendo descompensadas, elas podem piorar, fazendo com que a necessidade de correção aumente. Além disso, o uso inconsistente ou irregular de óculos pode causar desconforto, cansaço visual e até mesmo problemas mais graves, como glaucoma ou cataratas em casos severos. Portanto, é importante usar os óculos conforme prescrito pelo seu oftalmologista para manter a saúde de sua visão.

Tabela 2. Exemplo de resposta errada

Já nos experimentos com RAG o modelo indexador utilizado foi o e5-multilingual, com um tamanho de contexto de variável para análise de sensibilidade. O tamanho do contexto implica que os documentos do conjunto de dados foram divididos em pedaços iguais com X caracteres cada, onde X é o tamanho do contexto. Para cada pedaço de texto foi aplicado uma sobreposição de pedaços (*chunk overlap*) de 20% do total caracteres, ou seja, os pedaços de texto vizinhos compartilham 20% dos caracteres entre si, essa técnica é importante para que exista uma continuidade entre os pedaços de texto próximos. Apenas o melhor contexto recuperado pelo modelo foi utilizado para compor o prompt final.

Para o HYDE e *Rewrite-Retrieve-Read* foi utilizado o próprio modelo Mistral para realizar a geração do documento hipotético ou reescrever a pergunta utilizada para a recuperação. Os *prompts* para a geração foram criados baseados nos trabalhos originais com pequenas adaptações para a língua portuguesa.

O desempenho foi avaliado utilizando *Ragas*, com modelo avaliador *GPT-3.5* e indexador avaliador *text-embedding-ada-002*, modelo indexador de código fechado da OpenAI. As métricas escolhidas foram: fidelidade ao contexto, relevância da resposta e relevância do contexto.

3.7. Exemplos de respostas

Nas tabelas 1 e 2 podemos visualizar um exemplo de resposta correta e um exemplo de resposta incorreta. Ambas foram geradas utilizando *Rewrite-Retrieve-Read* com tamanho

de contexto 500 caracteres. A primeira resposta, possui algumas incoerências linguísticas mas, conceitualmente é acertada. Já no segundo caso, a resposta está conceitualmente errada, não existe relação entre não uso de óculos com o aumento das ametropias. O texto cita ainda alguns problemas válidos do não uso correto dos óculos como desconforto e cansaço visual, porém, novamente erra ao afirmar que glaucoma e catarata podem ter relação com o não uso de óculos de grau. O modelo acertou a segunda questão na configuração de 750 caracteres com a mesma técnica.

3.8. Resultados

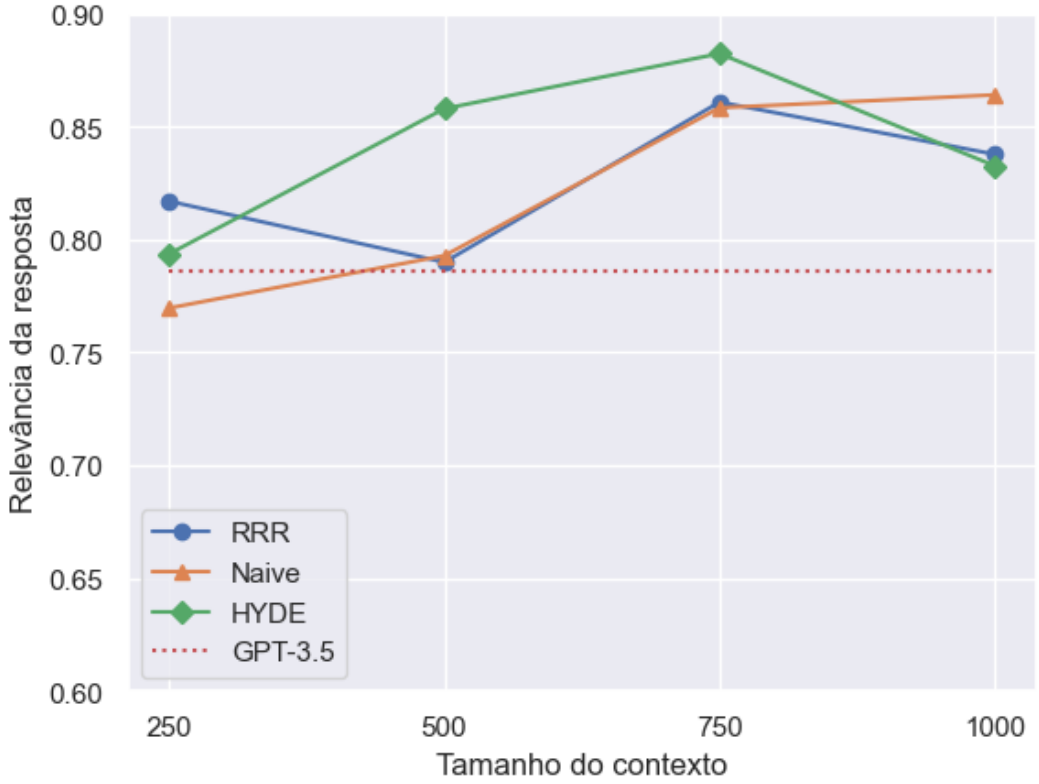

Figura 1. Relevância da resposta X Tamanho do contexto

A figura 1 ilustra o resultado da métrica relevância da resposta, comparando a linha de base com as três técnicas propostas. O gráfico demonstra que foi possível superar o baseline do *GPT-3.5* utilizando qualquer uma das três técnicas propostas. Destaca-se o tamanho de contexto 750 caracteres onde obteve-se a melhor performance para a relevância da resposta. A partir do tamanho de contexto 1000 visualiza-se uma queda na performance das respostas para HYDE e RRR, isso pode ser explicado por trabalhos como [Liu et al. 2024], onde é demonstrada a perda de performance ao utilizar contextos muito grandes. Devemos analisar a relevância da resposta em conjunto com as outras métricas para determinar o melhor intervalo de tamanho de contexto.

A figura 2 ilustra o resultado da métrica relevância do contexto, comparando as três técnicas propostas. Podemos visualizar a tendência decrescente da curva de relevância do contexto à medida que aumentamos o tamanho do contexto. Isso se explica pois a métrica aplicada penaliza os contextos que possuem alguma informação irrelevante para a resposta, algum ruído. É preciso tomar cuidado ao avaliar este gráfico pois,

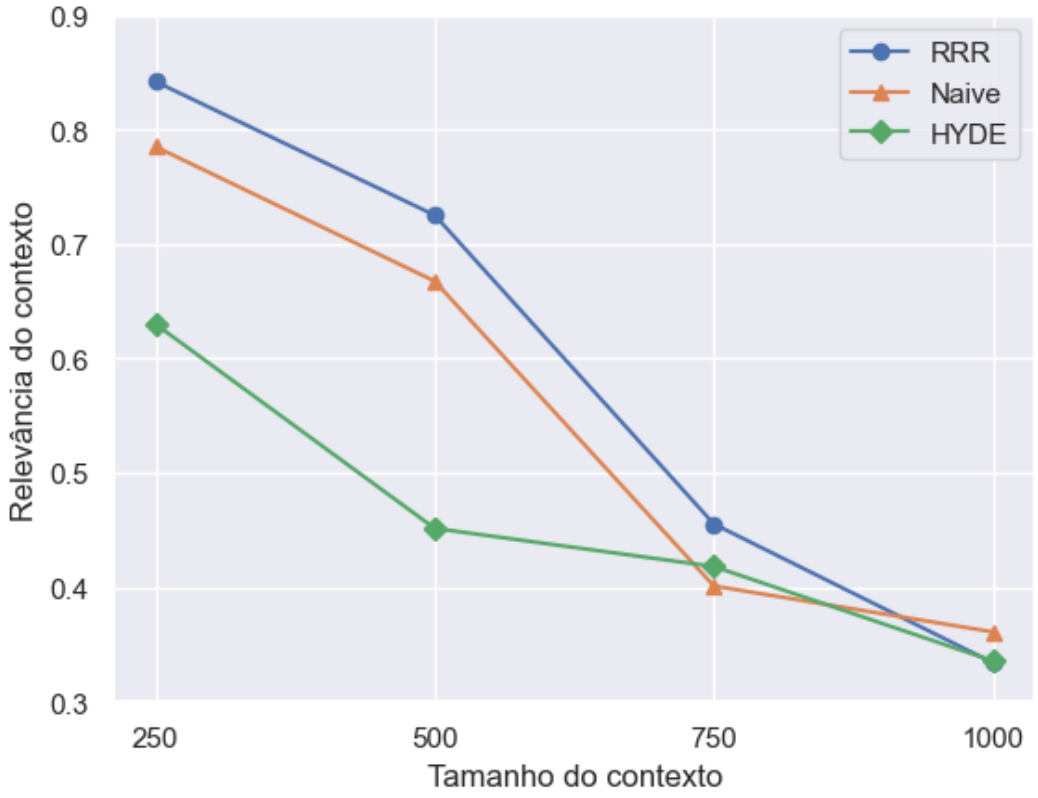

Figura 2. Relevância do contexto X Tamanho do contexto

um contexto pode ser relevante sem que seja completo, sem que contenha o máximo de informações possíveis para ajudar na composição da resposta. Em uma análise conjunta entre o resultado da relevância do contexto e da relevância da resposta, o destaque fica para o tamanho de contexto entre 500 e 750 caracteres, apresentando uma boa relevância de contexto, minimizando os ruídos, ao mesmo tempo que uma boa relevância da resposta, contendo o máximo de informações necessárias possíveis.

A figura 3 ilustra o resultado da métrica fidelidade ao contexto, comparando as três técnicas propostas. Nesta métrica visualiza-se um comportamento mais estável com uma pequena redução entre os tamanhos de contexto 500 e 750 caracteres. A redução desta métrica pode possuir dois significados: 1) o modelo não utiliza todo o contexto; 2) o modelo está utilizando informações paramétricas. Analisando a figura 3 em conjunto com a figura 2 podemos afirmar que essa redução trata-se do primeiro caso, onde o contexto está mais carregado de informações irrelevantes do que com o tamanho de contexto de 250.

A performance do HYDE não conseguiu superar o *naive* RAG na busca por contextos relevantes. Isto se explica analisando o próprio trabalho original do HYDE [Gao et al. 2023], em que a técnica não conseguiu obter a melhor performance no cenário multilingual, em especial quando o recuperador recebe um ajuste fino utilizando o conjunto de dados MS-MARCO [Nguyen et al. 2016] (conjunto de dados esse que foi utilizado no treinamento do e5-multilingual). Esse conjunto de dados ajuda o modelo a relacionar *query* ao contexto, reduzindo o problema de assimetria utilizando ajuste

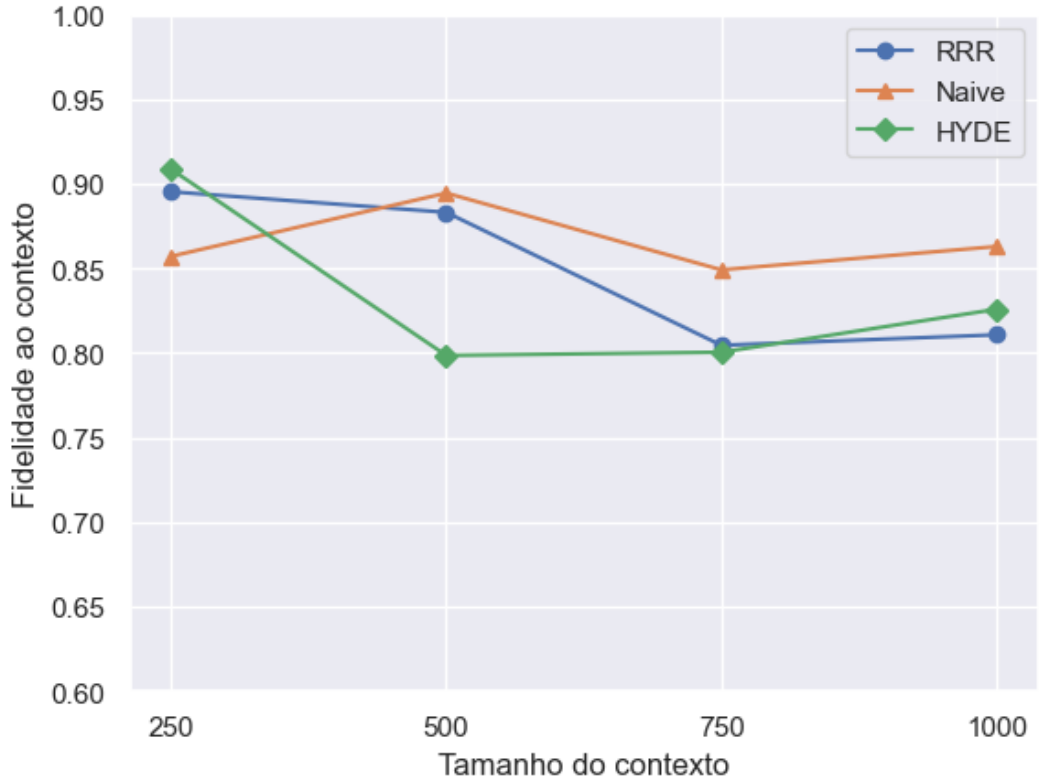

Figura 3. Fidelidade ao contexto X Tamanho do contexto

Tabela 3. Resultados para tamanho de contexto 750 caracteres RR: Relevância da resposta CR: Relevância do contexto FC: Fidelidade ao contexto

Modelo	RAG	RR	RC	FC
GPT-3.5-TURBO	-	0.7863	-	-
Mistral 7B	Naive	0.8584	0.4012	0.8492
Mistral 7B	HYDE	0.8824	0.4182	0.8005
Mistral 7B	RRR	0.8609	0.4554	0.8046

fino. O e5-multilingual possui, portanto, dois modos de busca: a busca simétrica e a busca assimétrica. A busca simétrica foi utilizada com a técnica HYDE uma vez que estamos comparando um documento hipotético com o documento real. Já a busca assimétrica foi utilizada nos outros experimentos. Seguindo portanto os resultados do trabalho [Gao et al. 2023] a busca assimétrica auxiliada por um modelo com ajuste fino performou melhor que a busca simétrica com HYDE.

Analisando as três figuras em conjunto é possível concluir que, o tamanho ideal de contexto para o caso analisado está entre 500 e 750 caracteres, apresentando boa relevância de resposta, de contexto e fidelidade ao contexto. O destaque ficou com a técnica *Rewrite-Retrieve-Read* que alcançou o melhor desempenho na relevância de contexto em quase todos os cenários.

As tabelas 3 e 4 detalham os resultados das métricas para os tamanhos de contexto 750 e 500 respectivamente. Podemos concluir que os resultados mais equilibrados

Tabela 4. Resultados para tamanho de contexto 500 caracteres RR: Relevância da resposta CR: Relevância do contexto FC: Fidelidade ao contexto

Modelo	RAG	RR	RC	FC
GPT-3.5-TURBO	-	0.7863	-	-
Mistral 7B	Naive	0.7929	0.6675	0.8946
Mistral 7B	HYDE	0.8580	0.4515	0.7986
Mistral 7B	RRR	0.7900	0.7252	0.8832

ficaram com o *naive* RAG e o RRR na configuração 500 caracteres. Porém, é notável a discrepância da relevância da resposta entre as duas abordagens. Essa discrepância aponta, portanto, para que a configuração mais otimizada esteja entre 750 e 500 caracteres. A dificuldade em escolher qual a melhor abordagem reside no *tradeoff* entre um contexto relevante e uma resposta completa; essa dificuldade foi abordada em trabalhos como: [Gao et al. 2024b] e ainda permanece como um problema em aberto.

Apenas a métrica relevância da resposta é possível ser mensurada para a linha de base do *ChatGPT*, uma vez que a linha de base foi construída sem informar contexto adicional ao modelo e portanto, as métricas de relevância do contexto e fidelidade ao contexto não podem ser aplicadas.

4. Conclusão e trabalhos futuros

Neste estudo, evidenciamos uma abordagem promissora na construção de *chatbots* para a área da saúde, destacando a viabilidade de alcançar desempenho igual ou superior aos modelos de mercado com um modelo menor em termos de parâmetros, sem ajuste fino, mediante o emprego de técnicas de RAG. Essa constatação ressalta a eficácia e acessibilidade dessas estratégias, permitindo a implementação de soluções eficientes, com aplicabilidade similar aos trabalhos relacionados na seção 2, mesmo com recursos limitados. Soluções locais, como a apresentada, vêm ao encontro do tipo de dado abordado, sendo que dado de saúde é considerado sensível. Portanto, a adoção de uma solução local satisfatória, que não envolva serviços terceirizados é preferível.

Os resultados mostraram que o tamanho ideal de contexto para essa aplicação está entre 500 e 750 caracteres, alcançando o melhor equilíbrio entre as três métricas aplicadas. A técnica de RAG *Rewrite-Retrieve-Read* performou significativamente melhor que as outras técnicas abordadas, em decorrência do cenário ser com perguntas no estilo informal, portanto é aconselhada no uso direto com o paciente.

A utilização de outro modelo de língua com alta acurácia, como o *ChatGPT*, no processo de avaliação é uma abordagem promissora com ciclos de avaliação mais rápidos. Assim, auxilia na otimização de técnicas de recuperação e geração de informação contribuindo para a precisão e relevância das respostas fornecidas pelo *chatbot*.

Além disso, destaca-se a importância de futuras pesquisas ampliarem tanto a quantidade quanto a qualidade das perguntas utilizadas na avaliação do *chatbot*, bem como validar os resultados por meio de avaliação humana, assegurando a confiabilidade e precisão do sistema desenvolvido. Avanços nesta área podem oferecer contribuições significativas para o campo da saúde pública, especialmente em especialidades como a oftalmologia, onde o acesso igualitário ao conhecimento e aos cuidados desde o pré até o pós tratamento

detém um impacto profundo no bem-estar e na qualidade de vida das pessoas.

Referências

Antaki, F., Touma, S., Milad, D., El-Khoury, J., and Duval, R. (2023). Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmology Science*, 3(4):100324.

Assi, L., Chamseddine, F., Ibrahim, P., Sabbagh, H., Rosman, L., Congdon, N., Evans, J., Ramke, J., Kuper, H., Burton, M. J., Ehrlich, J. R., and Swenor, B. K. (2021). A Global Assessment of Eye Health and Quality of Life: A Systematic Review of Systematic Reviews. *JAMA Ophthalmology*, 139(5):526–541.

Bernstein, I. A., Zhang, Y. V., Govil, D., Majid, I., Chang, R. T., Sun, Y., Shue, A., Chou, J. C., Schehlein, E., Christopher, K. L., Groth, S. L., Ludwig, C., and Wang, S. Y. (2023). Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Network Open*, 6(8):e2330320–e2330320.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Aletras, N. and De Clercq, O., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Gao, L., Ma, X., Lin, J., and Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Gao, M., Hu, X., Ruan, J., Pu, X., and Wan, X. (2024a). Llm-based nlg evaluation: Current status and challenges.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. (2024b). Retrieval-augmented generation for large language models: A survey.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Ma, X., Gong, Y., He, P., Zhao, H., and Duan, N. (2023). Query rewriting in retrieval-augmented large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset.

Organization, W. H. (2019). *World report on vision*. World Health Organization.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sulem, E., Abend, O., and Rappoport, A. (2018). BLEU is not suitable for the evaluation of text simplification. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2024). Text embeddings by weakly-supervised contrastive pre-training.

Zhao, H., Ling, Q., Pan, Y., Zhong, T., Hu, J.-Y., Yao, J., Xiao, F., Xiao, Z., Zhang, Y., Xu, S.-H., Wu, S.-N., Kang, M., Wu, Z., Liu, Z., Jiang, X., Liu, T., and Shao, Y. (2023). Ophtha-llama2: A large language model for ophthalmology.