Emanuel Borges Passinato

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

Goiânia

2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Emanuel Borges Passinato

Título do trabalho: Integração de Modelos de Linguagem e RAG na Criação de Chatbots Oftalmológicos

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 07/08/2024, às 20:09, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Emanuel Borges Passinato**, **Discente**, em 07/08/2024, às 20:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Emanuel Borges Passinato

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

Trabalho de conclusão de curso apresentado na Escola de Engenharia Elétrica, Mecânica e de Computação como requisito para a conclusão do curso de Engenharia de Computação e obtenção do título de Engenheiro de Computação.

Universidade Federal de Goiás – UFG

Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)

Orientador: Arlindo Rodrigues Galvão Filho

Coorientador: Walcy Santos Rezende Rios

Goiânia

2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Passinato, Emanuel Borges
Integração de Modelos de Linguagem e RAG na Criação de Chatbots Oftalmológicos [manuscrito] / Emanuel Borges Passinato. - 2024.
15 f.

Orientador: Prof. Dr. Arlindo Rodrigues Galvão Filho; co-orientador Walcy Santos Rezende Rios.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), Engenharia da Computação, Goiânia, 2024.

1. Chat bot. 2. Oftalmologia. 3. Retrieval augmented generation. 4. LLM. I. Galvão Filho, Arlindo Rodrigues, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Ao(s) quarto dia(s) do mês de junho do ano de 2024 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado “Integração de Modelos de Linguagem e RAG na Criação de Chatbots Oftalmológicos”, de autoria de Emanuel Borges Passinato, do curso de Engenharia de Computação, do(a) Escola de Engenharia Elétrica, Mecânica e de Computação da UFG. Os trabalhos foram instalados pelo(a) Prof. Dr. Arlindo Rodrigues Galvão Filho (INF) com a participação dos demais membros da Banca Examinadora: Prof. Dr. Thyago C. Marques (EMC) e Prof. Dr. Anderson da Silva Soares (INF). Após a apresentação, a banca examinadora realizou a arguição do(a) estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 9.5 , tendo sido o TCC considerado Aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 07/08/2024, às 19:58, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Thyago Carvalho Marques**, **Professor do Magistério Superior**, em 08/08/2024, às 08:04, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Anderson Da Silva Soares**, **Professor do Magistério Superior**, em 08/08/2024, às 09:35, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4728012** e o código CRC **5A5AE356**.

Referência: Processo nº 23070.017521/2024-53 SEI nº 4728012

Ao meus pais, Cairo César Borges Dias e Francisca Passinato Borges.

Agradecimentos

Gostaria de expressar meus agradecimentos à Universidade Federal de Goiás, que desde o início acolheu-me e guiou-me até o momento da conclusão deste curso de Engenharia de Computação, com muito respeito e dedicação por parte de todo o corpo docente e demais colaboradores.

Agradeço, também, ao meu professor orientador Arlindo Rodrigues Galvão Filho e ao coorientador Walcy Santos Rezende Rios pela dedicação e acompanhamento ao longo deste trabalho!

É com muita alegria que dedico essa conclusão de curso aos meus pais, que sempre me apoiaram e me ensinaram o valor da educação: Cairo César Borges Dias e Francisca Passinato Borges.

Finalmente, agradeço a Deus por todas as bênçãos a mim concedidas, em especial nas pessoas das irmãs franciscanas, do Colégio Santa Clara, setor Campinas, Goiânia-Go, através das quais fui agraciado com uma bolsa integral de estudos durante o ensino médio, possibilitando, assim, meu ingresso no ensino superior, onde agora finalizo este curso de graduação.

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

Emanuel B. Passinato
Centro de Competência Embrapii de Tecnologias Imersivas – AKCIT
Universidade Federal de Goiás
Goiânia – GO – Brasil
Email: emanuel.passinato@discente.ufg.br

Walcy S. R. Rios
Centro de Competência Embrapii de Tecnologias Imersivas – AKCIT
Universidade Federal de Goiás
Goiânia – GO – Brasil
Email: walcy.rios@discente.ufg.br

Arlindo R. Galvão Filho
Centro de Competência Embrapii de Tecnologias Imersivas – AKCIT
Universidade Federal de Goiás
Goiânia – GO – Brasil
Email: arlindo@ufg.br

***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.**

***Palavras Chave*—Chat bot, Oftalmologia, Retrieval augmented generation, LLM**

***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.**

***Index Terms*—Chat bot, Oftalmologia, Retrieval augmented generation, LLM**

1. Introdução

A oftalmologia desempenha um papel crucial e indispensável na manutenção e no aprimoramento da saúde ocular, com um impacto direto e profundo na qualidade de vida dos indivíduos [1]. Esta especialidade médica não se limita apenas ao tratamento de doenças que afetam os olhos, mas também abrange uma ampla gama de atividades voltadas para a preservação da visão e a promoção da saúde ocular em geral.

O olho humano é uma estrutura complexa e delicada, composta por diversos componentes interdependentes que trabalham em conjunto para permitir a percepção visual. Dada a sua complexidade, o olho requer cuidados especializados e monitoramento contínuo para prevenir e tratar uma vasta gama de condições que podem comprometer seriamente a visão e, consequentemente, a qualidade de vida. Estes cuidados incluem a detecção precoce de alterações visuais, o tratamento de doenças que vão desde condições relativamente benignas até patologias graves, e a gestão de problemas relacionados à visão que surgem com a idade ou devido a fatores genéticos e ambientais.

Além do tratamento das doenças oculares, a oftalmologia desempenha um papel essencial na prevenção de problemas visuais, o que envolve a realização de exames regulares e a orientação sobre práticas de saúde ocular. A especialidade também é fundamental na educação dos pacientes sobre a importância da saúde ocular e a necessidade de cuidados preventivos, como o uso de proteção adequada contra raios UV, a manutenção de uma dieta equilibrada e a prática de hábitos saudáveis que beneficiem a visão. Em suma, a oftalmologia não só busca tratar e curar, mas também promove um enfoque proativo para garantir uma visão saudável e de alta qualidade ao longo da vida.

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 [2] 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 [2].

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.

No entanto, os modelos de linguagem ainda enfrentam limitações notórias, especialmente quando são aplicados a tarefas que requerem um conhecimento aprofundado e especializado [**?**], como ocorre em tópicos relacionados à medicina. Essas tarefas que demandam um conhecimento intensivo frequentemente revelam as fraquezas desses modelos, que podem produzir respostas imprecisas ou até mesmo incorretas, levando a um efeito negativo sobre os usuários que dependem da precisão dessas informações.

Além disso, um desafio adicional significativo é a questão do sigilo dos dados ao utilizar serviços externos. Isso se torna particularmente relevante no contexto da área da saúde, onde os dados são naturalmente sensíveis e devem ser protegidos com rigor para garantir a privacidade e a segurança dos pacientes.

As recentes pesquisas na área [3], [4], exploram a utilização de modelos de mercado como o *ChatGPT* , ou modelos com ajuste fino como em [5]. 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) [6], 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: 1) A versão original proposta pelo trabalho [6], doravante referenciada como *naive* RAG; 2) *Hypothetical Document Embeddings* (HYDE) [7] e 3) *Rewrite-Retrieve-Read* [8].

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

2. Trabalhos Relacionados

Em [4] 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 [3] 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 [5] 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 privados 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 os quais 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.

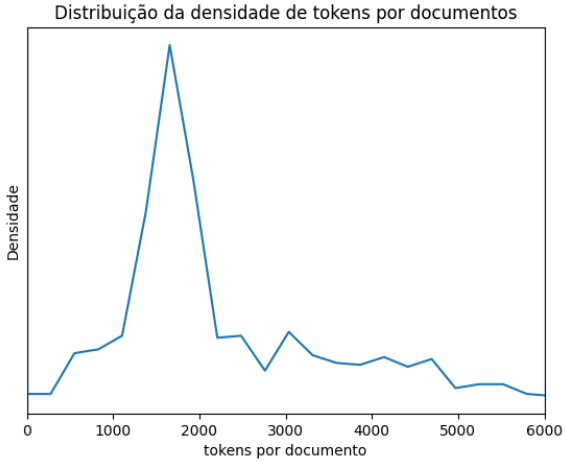

Figure 1. Distribuição da densidade de tokens por documentos

A figura 1 mostra a distribuição da quantidade de *tokens* para os documentos utilizados. A maioria dos documentos possui entre 1000 e 2000 tokens, sendo, portanto, considerados documentos de tamanho médio. É importante citar que o modelo gerador utilizado conseguiria receber um documento de ate 4 mil *tokens* como entrada, porém, como discutiremos a seguir, esta abordagem pode prejudicar a relevância do contexto fornecido.

Ao todo, os 714 documentos utilizados possuem 864.965 *tokens*. Alguns *outliers* com mais de 6 mil *tokens* foram retirados do gráfico para uma melhor visualização.

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 [9], pois diversos fatores a influenciam, como: subjetividade, fluidez, relevância, coerência, entre outros.

Métricas de avaliações tradicionais como *BLEU* [10] e *ROUGE* [11] 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 [12], 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. [13] 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 utilizadas 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* [13]

3.2.1. Fidelidade ao contexto. Para mensurar se uma resposta foi pautada ou não, no contexto, o RAGAS primeiramente utiliza um modelo gerador para extrair da resposta um conjunto de afirmações (S), ideias centrais que foram apresentadas como resposta. Posteriormente é pedido ao modelo que, de posse do contexto fornecido para geração, determinar as afirmações extraídas (S) que podem ser inferidas a partir do contexto (V).

O valor final, então, da métrica de fidelidade ao contexto pode ser calculado como:

$$FC = \frac{|V|}{|S|} \tag{1}$$

Onde V é a quantidade de afirmações que podem ser inferidas com base no contexto e S é a quantidade de afirmações feitas na resposta.

3.2.2. Relevância da resposta. Uma resposta relevante é considerada relevante, caso ela responda a pergunta diretamente e de uma maneira apropriada. Uma resposta pode ser relevante do ponto de vista da pergunta, mesmo que seja incorreta.

A métrica de relevância da reposta, portanto, não busca ponderar sobre a factualidade da resposta, mas se a resposta é completa, baseada na pergunta e não possui informações redundantes.

Para calcular a relevância da resposta, utilizamos um modelo gerador para criar perguntas à partir da resposta a ser analisada (qi). Para cada pergunta em potencial gerada, utilizamos um modelo indexador para calcular a similaridade com a pergunta original. O valor final da relevância da resposta é, então, obtido como:

$$RR = \frac{1}{n}\sum_{1}^{n} \text{coss\_sim}(qi, q) \tag{2}$$

Portanto, essa métrica avalia o quão próximas as perguntas geradas, a partir da resposta, estão da pergunta original.

3.2.3. Relevância do contexto. Um contexto pode ser considerado relevante quando tem apenas informações que são úteis para responder a pergunta. Esta métrica penaliza, assim, contextos com informações irrelevantes ou redundantes.

Podemos estimar a relevância do contexto, extraindo dele o subconjunto de sentenças que contém informações

relevantes para responder a pergunta, comparando posteriormente a quantidade de sentenças relevantes com a quantidade total de sentenças do contexto. Definindo a equação como:

$$RR = \frac{\text{quantidade de sentenças relevantes}}{\text{quantidade total de sentenças}} \quad (3)$$

Portanto, valorizamos apenas contextos que possuem informações relevantes para responder a dada pergunta.

3.3. *Retrieval Augmented Generation* - RAG

Neste estudo foram exploradas três abordagens de RAG: 1) *naive* RAG, 2) *hypothetical document embeddings* (HYDE) e 3) *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 entre os vetores latentes da pergunta e os dos documentos fornecidos como base de dados.

Os modelos indexadores desempenham um papel crucial na organização e recuperação de informações, recebendo um texto como entrada e produzindo, como saída, um vetor que representa as características e a essência desse texto. Esse vetor serve como uma representação compacta e matemática do conteúdo textual, facilitando a comparação e a busca por informações relevantes. A estrutura de dados resultante é uma associação chave-valor, onde a "chave" é o vetor, sendo este um elemento numérico multidimensional que encapsula as propriedades semânticas e contextuais do texto original, e o "valor" é o próprio texto original.

Essa estrutura de chave-valor permite uma indexação mais adequada, pois a busca por documentos pode ser realizada com base nas características vetoriais em vez de correspondências exatas de texto. Quando um novo texto é inserido no sistema, ele é convertido em um vetor que pode ser rapidamente comparado com os vetores já existentes na base de dados, facilitando a recuperação de documentos relevantes com alta precisão. Esse método não só melhora a eficiência na busca e na organização de grandes volumes de dados, mas também permite a manipulação e análise avançada das informações com base nas características vetoriais dos textos. Assim, a abordagem chave-valor, com vetores como chaves e textos como valores, proporciona uma maneira poderosa de estruturar e acessar dados textuais em sistemas de recuperação de informação e análise de dados.

O cálculo de similaridade entre documentos pode ser realizado de várias maneiras, dependendo do contexto e das necessidades específicas da análise. Uma das abordagens mais comuns e amplamente utilizada, que foi adotada neste estudo, é a similaridade do cosseno. Esta técnica é particularmente eficaz para medir a similaridade entre vetores em um espaço vetorial, como aqueles gerados por um indexador.

A medida de similaridade do cosseno baseia-se no cosseno do ângulo formado entre dois vetores no espaço vetorial. Esses vetores, que são o resultado do processo de indexação, representam características ou atributos dos documentos em análise. A similaridade do cosseno quantifica o grau de alinhamento entre os vetores, fornecendo uma métrica que indica o quanto os documentos são similares entre si. A fórmula utilizada para calcular a similaridade do cosseno é a seguinte:

$$cos\Theta = \frac{A \cdot B}{\|A\| \, \|B\|} \quad (4)$$

Onde:

$$\|A\| = \sqrt{\sum_{1}^{n} A_i^2} \quad (5)$$

E ainda:

$$\|B\| = \sqrt{\sum_{1}^{n} B_i^2} \quad (6)$$

Na prática, vetores que apontam para direções próximas, apresentam o resultado do cosseno próximo de um, vetores que não possuem nenhuma relação apresentam resultado zero e vetores que possuem sentidos opostos apresentam resultados próximos a -1.

Portanto, após calcular o cosseno, os vetores cujo resultado for mais próximo de um são os mais similares.

3.3.1. *Naive RAG*. Os documentos de maior relevância para responder à consulta do usuário são identificados com base na similaridade entre vetores. O processo começa com a entrada do usuário, conhecida como *query*. O modelo indexador converte essa entrada em um vetor de características, que é, então, comparado com as chaves (*keys*) na base vetorial. Essas chaves são vetores representativos de documentos armazenados na base de dados.

Após essa etapa inicial de comparação, os documentos cujos vetores são mais semelhantes ao vetor da entrada do usuário, são recuperados. Esses documentos selecionados são, então, combinados com a entrada original do usuário, para criar um novo *prompt*. Esse novo *prompt* incorpora tanto a consulta do usuário quanto as informações relevantes extraídas dos documentos semelhantes.

O *prompt* resultante é, desse modo, enviado ao modelo de linguagem para a geração da resposta. Este método é projetado para fornecer ao modelo gerador um contexto mais rico e específico, garantindo que ele tenha acesso às informações necessárias para formular uma resposta precisa e relevante. Ao reduzir a probabilidade de alucinações, ou seja, respostas incorretas ou imprecisas e ao aumentar a relevância da resposta gerada, o processo melhora significativamente a qualidade e a utilidade das respostas fornecidas ao usuário.

3.3.2. *Hypothetical Document Embeddings (HYDE)*. HYDE [7] é uma técnica avançada de Recuperação e Geração de Respostas (RAG) que se destina a superar um problema recorrente na abordagem tradicional do *naive RAG*: a assimetria entre a pergunta e o documento de contexto utilizado para formulá-la.

Essa assimetria pode se manifestar de várias maneiras, com as principais causas sendo: a diferença de tamanho entre as sentenças, a discrepância nos estilos de escrita e a presença de ruídos na formulação da pergunta. O impacto dessa assimetria resulta em uma imprecisão na recuperação dos documentos mais relevantes e pertinentes para a resposta desejada.

Para mitigar esses problemas, a abordagem proposta pelo HYDE consiste em ajustar o texto de entrada fornecido pelo usuário para aproximá-lo dos documentos previamente indexados, gerando um documento hipotético. Esse documento hipotético é formulado com base na entrada do usuário, buscando ser mais congruente com o conteúdo esperado do que se utilizasse a entrada bruta diretamente.

Embora o documento hipotético possa ainda apresentar algumas alucinações ou imprecisões, ele oferece uma aproximação mais relevante em relação ao documento esperado. Com esse texto hipotético em mãos, o *retriever* é capaz de identificar e recuperar o documento, ou um conjunto de documentos, mais semelhantes para compor o *prompt* final utilizado na geração da resposta. Esse processo aprimorado visa otimizar a precisão e a relevância das respostas geradas, superando a limitação de assimetria entre a pergunta e os documentos de contexto.

3.3.3. *Rewrite-Retrieve-Read* (RRR). A ideia central por trás do método *Rewrite-Retrieve-Read* [8] é semelhante à do HYDE, no sentido de que ambos envolvem a reescrita do texto original para facilitar a busca de documentos relevantes. No entanto, as diferenças entre os dois métodos residem na abordagem adotada para a reescrita do texto.

No *Rewrite-Retrieve-Read*, a abordagem envolve a extração de uma ou mais perguntas independentes, a partir da pergunta original. Essas perguntas reescritas são então usadas para realizar buscas na base vetorial de conhecimento.

Esse processo é projetado para minimizar possíveis ruídos e ambiguidades presentes na pergunta original, além de segmentar a questão quando ela abrange múltiplos temas. Ao fazer isso, o método permite que a busca na base de conhecimento seja feita de maneira mais precisa e focada, facilitando a obtenção de informações mais relevantes e específicas. Assim, a reescrita do texto tem o objetivo de separar temas distintos e melhorar a eficácia da busca ao tratar cada aspecto da pergunta original de forma independente.

3.4. Modelo gerador: Mistral

O modelo escolhido para a geração das respostas foi o Mistral [14], um modelo de linguagem com 7 bilhões de parâmetros, desenvolvido com uma combinação de inovação tecnológica e foco estratégico em desempenho e eficiência. Este modelo destaca-se por empregar técnicas avançadas como a *grouped-query attention* (GQA), que visa otimizar o processo de inferência ao acelerar a capacidade do modelo de processar informações e gerar respostas de forma mais eficiente. Além disso, o Mistral utiliza a *sliding window attention* (SWA), uma técnica projetada para lidar com sequências de entrada de comprimento variável, de maneira mais flexível e eficaz. Esta abordagem reduz o custo associado à inferência e melhora a capacidade do modelo de gerenciar e processar dados complexos com menor latência.

A eficiência e o desempenho do Mistral foram comprovados em comparação com outros modelos de linguagem. Em testes de desempenho, o Mistral mostrou-se superior a modelos maiores, como o modelo 13B (Llama 2), e, em alguns cenários, até mesmo superou o modelo 34B (Llama 1), que possui um número maior de parâmetros. Essas comparações evidenciam que, apesar de seu tamanho relativamente menor, o Mistral oferece uma qualidade de resposta e eficiência que rivaliza e, em alguns casos, supera a dos modelos maiores. Esse desempenho superior não só confirma a eficácia das técnicas empregadas pelo Mistral, mas também evidencia sua relevância e potencial como uma solução eficaz para aplicações que exigem um equilíbrio entre capacidade computacional e eficiência operacional.

3.5. Modelo indexador: e5-multilingual

E5 [15] 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.

A versão deste modelo utilizada foi a multilíngua *base*, apresentada no trabalho [16]. As capacidades multilíngues do e5 foram avaliadas em dois grandes *benchmarks* para modelos indexadores MIRACL [17] e Bitext mining [18].

O desempenho do modelo nesses *benchmarks* se comparou, ou até superou modelos que realizaram treinamentos nos conjuntos de dados de treino dos *benchmarks* avaliados, apesar do e5 não ter utilizado tais conjuntos de dados. Esse resultado reforça o poder deste modelo para tarefas *zero shot* como a proposta neste trabalho.

Outra vantagem desse modelo é a capacidade de alternar entre busca simétrica e assimétrica. Discutimos sobre assimetria quando falamos sobre o HYDE e como ela impacta negativamente a etapa de recuperação. O e5 possui treinamento para realizar tanto busca simétrica, como busca assimétrica.

Para os casos de Naive RAG e *Rewrite-Retrieve-Read* utilizamos a busca assimétrica, pois estamos relacionando pergunta com documento. Para o HYDE utilizamos a busca simétrica, pois estamos relacionando documento hipotético com documento.

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

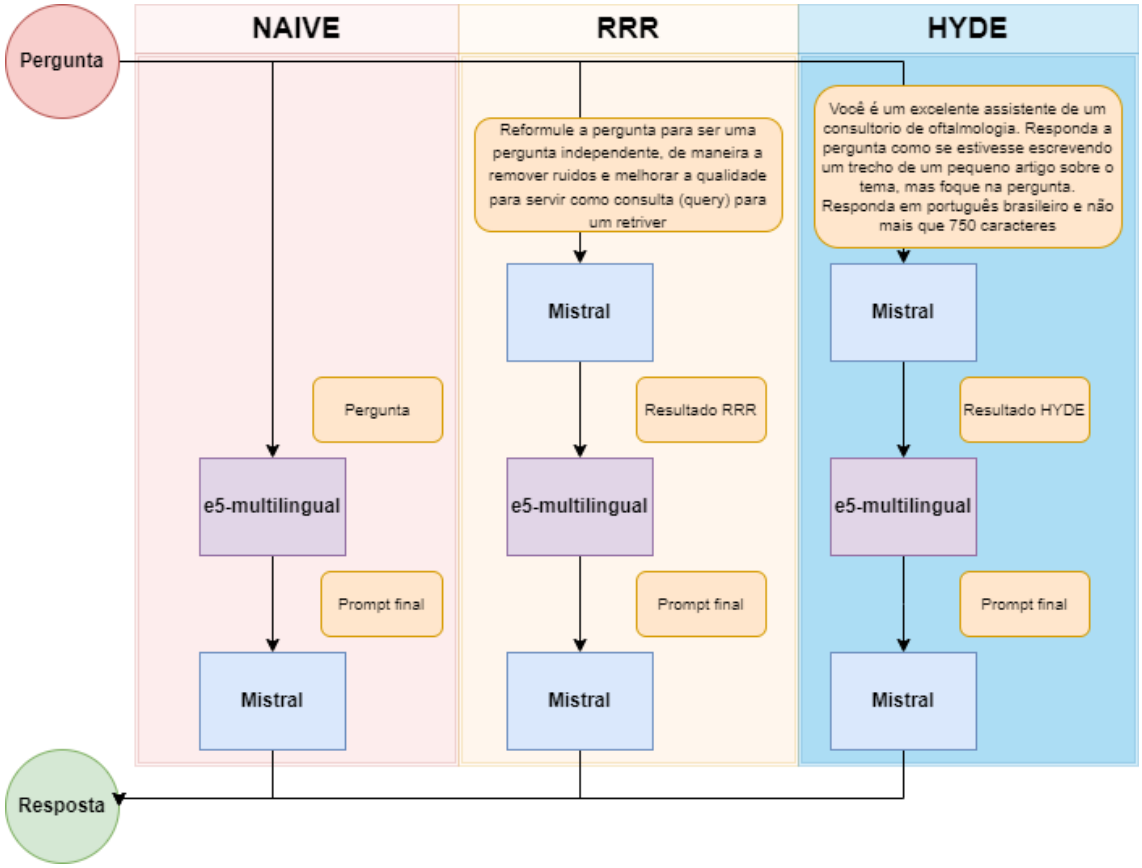

Figure 2. Diagrama representando a montagem experimental. Naive: Naive RAG; RRR: *Rewrite-Retrieve-Read*; HYDE: Hypothetical Document Embeddings.

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.

A figura 2 ilustra, em forma de diagrama, como foram organizados os experimentos para as três abordagens com RAG. Nela podemos observar as diferenças entre as três técnicas, bem como o fluxo geral dos dados.

Recebendo a pergunta como entrada, recuperamos o documento mais similar, utilizando o modelo indexador e posteriormente compomos o *prompt* final para o modelo gerador. A diferença entre as abordagens está justamente no texto que é fornecido para o modelo indexador.

Nos experimentos com RAG o modelo indexador utilizado foi o e5-multilingual, com um tamanho de contexto de variável entre 250, 500, 750 e 1000 caracteres, para análise de sensibilidade.

O tamanho do contexto implica que os documentos do conjunto de dados foram divididos em pedaços (*chunks*) 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 de 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.

3.6.1. HYDE. Para a aplicação da técnica HYDE, o *prompt* utilizado para a criação da resposta hipotética foi formulado da seguinte maneira:

> ”Você é um excelente assistente de um consultório de oftalmologia. Responda à pergunta como se estivesse escrevendo um trecho de um pequeno artigo sobre o tema, mas foque especificamente na pergunta. Responda em português brasileiro e limite-se a no máximo X caracteres.”

Esse *prompt* foi elaborado com base nos exemplos descritos no trabalho original e tem o objetivo primordial de orientar o modelo para gerar uma resposta hipotética que mitigue o problema de assimetria entre a pergunta e os documentos de contexto. A ideia é que, ao seguir essas diretrizes, o modelo seja capaz de produzir respostas mais precisas e relevantes.

É importante observar que o *prompt* especifica três aspectos cruciais para a geração da resposta: a adequação da língua (português brasileiro), o gênero textual desejado (artigos curtos), e o limite máximo de caracteres permitido (X caracteres). Essas especificações ajudam a garantir que

a resposta gerada esteja no formato correto e dentro das expectativas estabelecidas para a tarefa.

Além disso, em cada avaliação de caracteres, o valor de X é ajustado conforme a quantidade específica de caracteres que queremos testar. Essa abordagem permite uma adaptação precisa da resposta em função do comprimento do contexto, ajudando a calibrar a técnica para obter o melhor equilíbrio entre detalhamento e concisão.

3.6.2. Rewrite-Retrieve-Read. O *prompt* utilizado para reescrever a pergunta foi formulado da seguinte maneira:

> "Reformule a pergunta para que ela se torne uma pergunta independente, de modo a remover quaisquer ruídos e melhorar sua qualidade para que possa servir como uma consulta (*query*) eficaz para um *retriever*."

.

Este *prompt* foi cuidadosamente criado com base no exemplo fornecido no trabalho original e tem como objetivo principal orientar o modelo para gerar uma entrada mais clara e objetiva para o indexador. A ideia é que a pergunta reformulada seja livre de ambiguidades e ruídos que possam comprometer a eficiência do processo de recuperação de informações.

A reformulação pretendida visa aprimorar a qualidade da pergunta, transformando-a em uma consulta que seja bem definida e precisa. Isso é essencial para garantir que o *retriever* possa processar a consulta de forma eficiente e eficaz, resultando na recuperação dos documentos mais relevantes e pertinentes, para responder à pergunta original.

Em outras palavras, o *prompt* tem a intenção de transformar a pergunta inicial, potencialmente vaga ou mal formulada, em uma consulta que maximiza a precisão e a relevância das respostas geradas pelo sistema. Ao remover ruídos e garantir clareza o *prompt* ajuda a melhorar a eficácia geral do sistema de recuperação de informações, facilitando a obtenção de respostas mais úteis e informativas.

Tanto para o HYDE como para o *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 gerador avaliador *GPT-3.5* e indexador avaliador *text-embedding-ada-002*, modelo indexador de código fechado da OpenAI. As métricas utilizadas foram: fidelidade ao contexto, relevância da resposta e relevância do contexto.

3.7. Exemplos de respostas

Nas tabelas 1 e 2, oferecemos exemplos de uma resposta correta e uma resposta incorreta, ambas geradas utilizando a técnica de *Rewrite-Retrieve-Read* com um contexto de 500 caracteres. Esses exemplos servem para evidenciar as nuances e a precisão do processo de geração de respostas, permitindo uma análise detalhada do desempenho do modelo.

TABLE 1. EXEMPLO DE RESPOSTA CORRETA

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

TABLE 2. EXEMPLO DE RESPOSTA ERRADA

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.

A primeira resposta exemplifica um caso em que a técnica de geração de respostas é bem-sucedida em termos conceituais, apesar de algumas incoerências linguísticas. Isso implica que, enquanto o texto pode apresentar erros na sua construção ou em sua clareza, a informação essencial transmitida está correta e alinhada com o conhecimento esperado sobre o tema. O modelo conseguiu capturar e articular adequadamente o conteúdo necessário para fornecer uma resposta precisa.

Por outro lado, a segunda resposta ilustra um caso em que a técnica falha na precisão conceitual. Neste exemplo, o modelo não estabelece uma relação válida entre a ausência de uso de óculos e o aumento das ametropias, o que é um erro grave, já que essa correlação não é suportada pela literatura médica.

A resposta também menciona problemas reais associados ao uso inadequado dos óculos, como desconforto e cansaço visual, mas comete um engano ao relacionar o não uso de óculos a condições sérias como glaucoma e catarata, o que não é corroborado por evidências científicas.

A análise revela que o modelo, na configuração com 750 caracteres, conseguiu corrigir esses erros conceituais, mostrando a importância de um contexto mais amplo para aprimorar a precisão das respostas geradas. Essa melhoria demonstra que o aumento do tamanho do contexto pode impactar a qualidade e a exatidão das respostas, ajudando a reduzir falhas conceituais e a gerar informações mais corretas e relevantes.

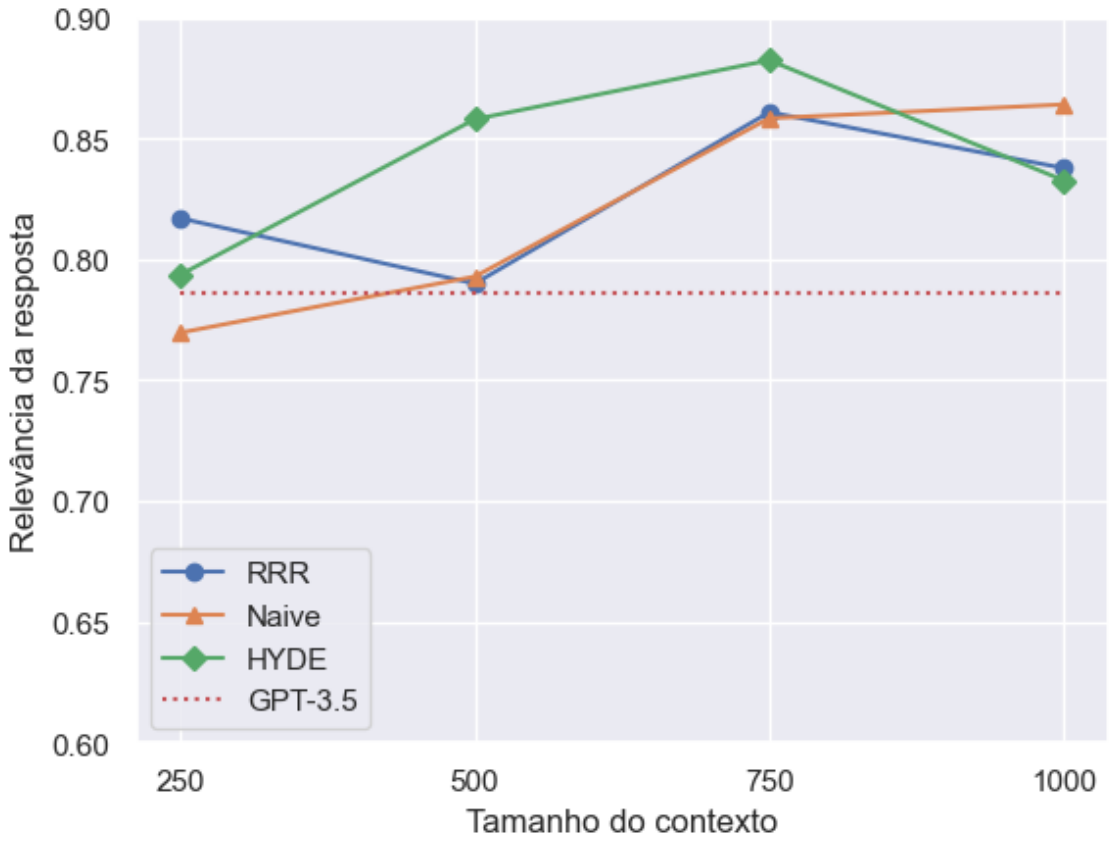

Figure 3. Relevância da resposta X Tamanho do contexto.

3.8. Resultados

A figura 3 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.

A figura 3 ilustra de forma detalhada o resultado obtido pela métrica de relevância da resposta, apresentando uma comparação entre a linha de base e as três técnicas propostas. O gráfico resultante evidencia que foi possível superar o desempenho da linha de base do *GPT-3.5* ao aplicar qualquer uma das três técnicas sugeridas, o que indica uma melhoria significativa em relação ao método convencional.

O gráfico revela que o tamanho de contexto de 750 caracteres proporcionou a melhor performance em termos de relevância da resposta. Esse resultado destaca a eficácia do ajuste de contexto para otimizar a qualidade das respostas geradas, com um desempenho superior em comparação aos outros tamanhos de contexto avaliados.

Contudo, ao analisar o tamanho de contexto de 1000 caracteres, observa-se uma queda na performance das respostas para as técnicas HYDE e RRR. Esse declínio na eficácia pode ser explicado por estudos anteriores, como o trabalho de [19], que demonstram que o uso de contextos excessivamente grandes pode resultar em perda de performance.

Tais trabalhos indicam que, embora contextos maiores possam inicialmente parecer benéficos, eles podem, na prática, introduzir ruídos e informações irrelevantes que comprometem a qualidade das respostas geradas.

É fundamental considerar a relevância da resposta em conjunto com outras métricas de avaliação, ao determinar o intervalo ideal de tamanho de contexto. Esse enfoque abrangente permitirá uma análise mais precisa e a escolha do tamanho de contexto que ofereça o melhor equilíbrio entre relevância e performance, otimizando assim a eficácia geral do sistema de recuperação e geração de respostas.

A figura 4 oferece uma visão detalhada dos resultados obtidos pela métrica de relevância do contexto, apresentando uma comparação entre as três técnicas propostas. O gráfico ilustra uma tendência clara de declínio na relevância do contexto à medida que o tamanho do contexto aumenta.

Esse comportamento pode ser compreendido pelo fato de que a métrica de relevância penaliza contextos que contêm informações irrelevantes ou ruídos, que não contribuem diretamente para a formulação de uma resposta precisa.

Esse fenômeno ocorre porque contextos mais extensos têm uma maior probabilidade de incluir informações desnecessárias, ou menos pertinentes, que acabam compro-

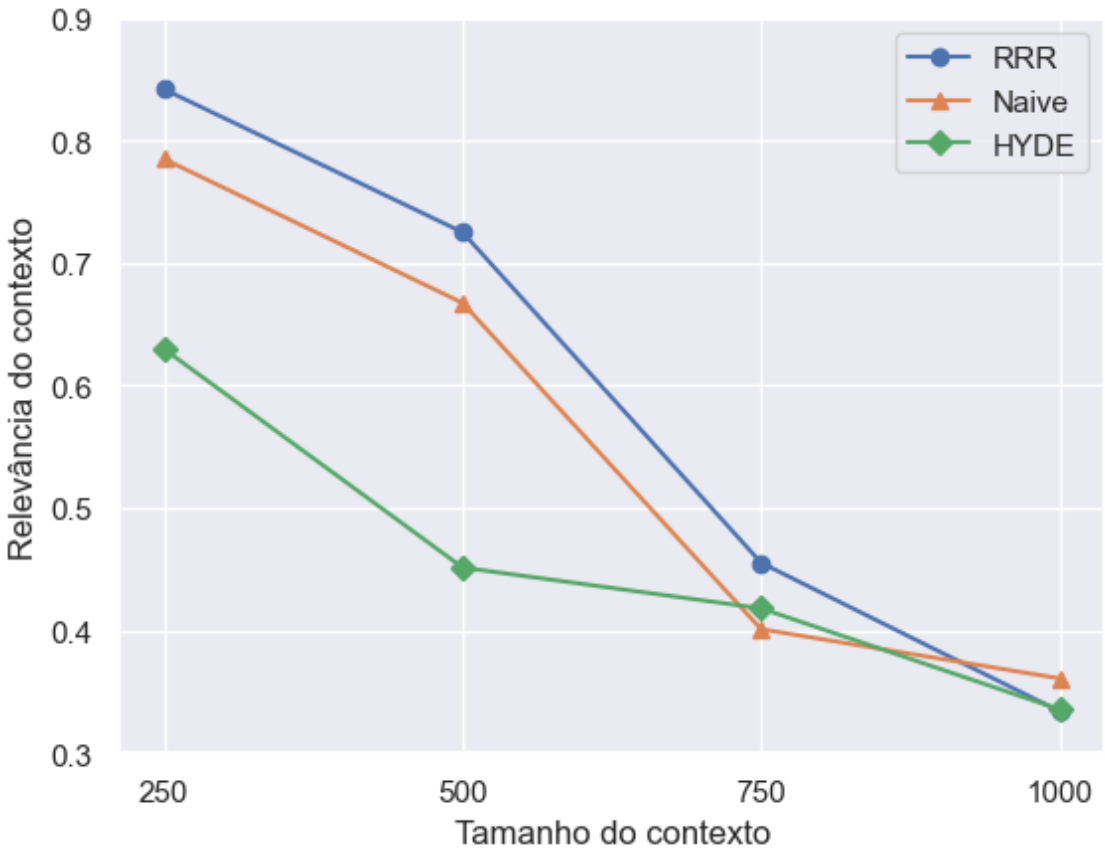

Figure 4. Relevância do contexto X Tamanho do contexto

metendo a qualidade da relevância do contexto.

A presença de tais informações irrelevantes pode dificultar a identificação dos dados verdadeiramente úteis para a resposta, resultando em uma pontuação inferior na métrica de relevância.

No entanto, é crucial interpretar este gráfico com cautela. Um contexto pode ser relevante sem que seja completo, sem que contenha o máximo de informações possíveis para ajudar na composição da resposta.

É possível que um contexto contenha uma quantidade significativa de informações relevantes, mas não seja completamente adequado se estiver sobrecarregado com dados que não são diretamente úteis para a resposta.

Ao realizar uma análise integrada dos resultados da relevância do contexto em conjunto com a relevância da resposta, observa-se que o tamanho de contexto entre 500 e 750 caracteres se destaca. Este intervalo apresenta uma combinação ideal, oferecendo uma boa relevância de contexto ao minimizar a presença de ruídos e informações irrelevantes. Simultaneamente, mantém uma alta relevância da resposta ao incluir a quantidade adequada de informações necessárias para gerar respostas precisas e completas.

A figura 5 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 5 em conjunto com a figura 4 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 [7], em que a técnica não conseguiu obter a melhor performance no cenário multilíngua, em especial quando o recuperador recebe um ajuste fino utilizando o conjunto de dados MS-MARCO [20] (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 fino.

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 [7] a busca

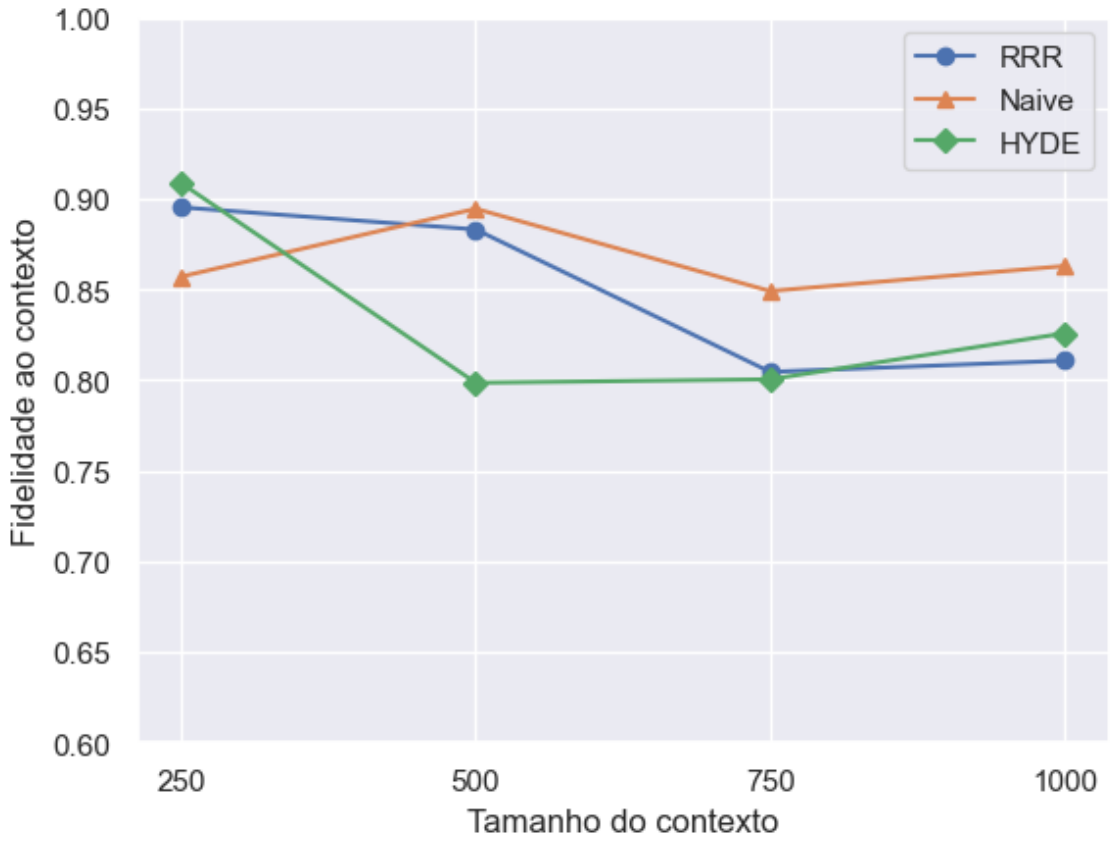

Figure 5. Fidelidade ao contexto X Tamanho do contexto

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 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: [21] 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.

TABLE 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

TABLE 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

4. Conclusão

Neste estudo, destacamos uma abordagem que consideramos promissora na construção de *chatbots* voltados para a área da saúde, evidenciando a viabilidade de alcançar um desempenho igual ou até superior ao dos modelos de mercado, utilizando um modelo com menos parâmetros e sem a necessidade de ajuste fino, graças à aplicação de técnicas de *retrieval augmented generation* (RAG).

Esta abordagem não só demonstra a eficácia das técnicas empregadas mas, também, evidencia a possibilidade de atingir um desempenho elevado com recursos limitados, sem comprometer a qualidade.

Essa constatação ressalta a eficácia e a acessibilidade das estratégias adotadas, permitindo a implementação de soluções eficientes que têm aplicabilidade comparável à dos trabalhos relacionados descritos na seção 2, mesmo em cenários com recursos limitados. A capacidade de obter resultados significativos sem a necessidade de grandes investimentos em modelos complexos representa uma vantagem considerável, oferecendo uma alternativa viável e econômica para o desenvolvimento de *chatbots* na área da saúde.

Além disso, a implementação de soluções locais, como a apresentada neste estudo, é particularmente relevante considerando a natureza sensível dos dados de saúde. Dados desta natureza requerem um manejo cuidadoso e a proteção da privacidade dos usuários. Portanto, adotar uma solução local satisfatória, que não dependa de serviços terceirizados, é preferível e mais alinhada com as melhores práticas de segurança e privacidade.

A utilização dessa abordagem não apenas reduz os riscos associados ao compartilhamento de dados sensíveis, mas também reforça a confiança dos usuários na proteção de suas informações pessoais.

Os resultados obtidos demonstraram que o tamanho ideal do contexto para a aplicação em questão está entre 500 e 750 caracteres. Esse intervalo foi identificado como o ponto de equilíbrio mais eficaz, levando em consideração as três métricas de avaliação aplicadas no estudo.

Essa faixa de comprimento de contexto oferece uma combinação ideal de detalhe e concisão, permitindo que o modelo compreenda e responda de maneira precisa e relevante às perguntas, sem sobrecarregar o sistema com informações excessivas ou insuficientes.

A técnica de RAG *Rewrite-Retrieve-Read*, se destacou em relação às outras técnicas avaliadas. A superioridade desta abordagem pode ser atribuída ao fato de que o cenário em questão envolve perguntas de estilo informal, que exigem uma adaptação mais flexível e contextualizada das respostas.

O método *Rewrite-Retrieve-Read* demonstrou uma capacidade notável para lidar com essas particularidades, oferecendo respostas mais precisas e relevantes para o usuário final. Portanto, recomenda-se a utilização desta técnica quando se trata de interações diretas com pacientes, dada sua eficácia em gerenciar a natureza informal e frequentemente variável das perguntas e respostas neste contexto.

A incorporação de um modelo de linguagem, como o *ChatGPT*, no processo de avaliação representa uma abordagem promissora para acelerar os ciclos de avaliação e aprimorar a eficiência geral do sistema. A utilização do *ChatGPT* permite avaliações mais rápidas e detalhadas, facilitando a otimização das técnicas de recuperação e geração de informações. Isso resulta em respostas mais precisas e relevantes fornecidas pelo *chatbot*, atendendo melhor às necessidades dos usuários.

Para implementar essa abordagem, foi utilizado o *framework RAGAS*, que oferece uma estrutura eficiente para avaliar o desempenho com diferentes modelos de linguagem. O uso do RAGAS possibilitou uma análise detalhada e ajustes mais eficazes nas técnicas de recuperação e geração de informações. Dessa forma, a combinação do *ChatGPT* com o *framework RAGAS* não só melhora a qualidade das respostas do *chatbot*, mas também torna o processo de avaliação mais ágil e eficiente.

Destaca-se, ainda, a importância de futuras pesquisas no sentido de ampliar, 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.

Apêndice

Apêndice A.

Todas as tabelas contendo as métricas extraídas estão disponíveis nesta seção do Apêndice, organizadas de maneira a proporcionar uma visão detalhada e sistemática dos dados coletados. As tabelas foram divididas conforme o tamanho do contexto utilizado em cada teste, o que facilita a comparação entre as técnicas abordadas, de forma mais eficiente e transparente.

Além disso, a organização das tabelas por tamanho de contexto permite uma comparação direta e clara entre as técnicas, revelando como cada uma delas se comporta em cenários variados. Esta disposição facilita a identificação de padrões e tendências, ajudando a entender melhor como diferentes estratégias influenciam os resultados.

Ao estruturar os dados dessa forma, a seção do Apêndice não apenas oferece uma visão abrangente das métricas, mas também proporciona uma ferramenta valiosa para uma análise crítica e comparativa das técnicas discutidas no estudo. Com isso, os leitores podem realizar uma avaliação mais informada das abordagens apresentadas e das suas respectivas eficácias.

TABLE 5. RESULTADOS PARA TAMANHO DE CONTEXTO 250 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	0.7697	0.7850	0.8946
Mistral 7B	HYDE	0.7937	0.6299	**0.9087**
Mistral 7B	RRR	**0.8170**	**0.8419**	0.8954

TABLE 6. 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

TABLE 7. 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

TABLE 8. RESULTADOS PARA TAMANHO DE CONTEXTO 1000 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	**0.8641**	**0.3610**	**0.8630**
Mistral 7B	HYDE	0.8327	0.3357	0.8259
Mistral 7B	RRR	0.8378	0.3341	0.8108

Apêndice B.

Todas as tabelas contendo as métricas extraídas estão disponíveis nesta seção do Apêndice, organizadas de maneira a proporcionar uma visão detalhada e sistemática dos dados coletados. As tabelas foram divididas conforme as técnicas utilizadas em cada teste, o que facilita a comparação entre os tamanhos de contexto abordados, de forma mais eficiente e transparente.

Além disso, a organização das tabelas permite uma comparação direta e clara entre os tamanhos de contextos, revelando como cada um deles se comporta em cenários variados. Esta disposição facilita a identificação de padrões e tendências, ajudando a entender melhor como diferentes tamanhos de contexto influenciam os resultados.

Ao estruturar os dados dessa forma, a seção do Apêndice não apenas oferece uma visão abrangente das métricas, mas também proporciona uma ferramenta valiosa para uma análise crítica e comparativa dos tamanhos de contextos discutidos no estudo. Com isso, os leitores podem realizar uma avaliação mais informada das abordagens apresentadas e das suas respectivas eficácias.

TABLE 9. RESULTADOS PARA TÉCNICA *Naive RAG*

Tamanho Contexto	RR	RC	FC
250	0.7697	**0.7850**	0.8572
500	0.7929	0.6675	**0.8946**
750	0.8584	0.4012	0.8492
1000	**0.8641**	0.3610	0.8630

TABLE 10. RESULTADOS PARA TÉCNICA *HYDE*

Tamanho Contexto	RR	RC	FC
250	0.7937	**0.6299**	**0.9087**
500	0.8580	0.4515	0.7986
750	**0.8824**	0.4182	0.8005
1000	0.8327	0.3357	0.8259

TABLE 11. RESULTADOS PARA TÉCNICA *RRR*

Tamanho Contexto	RR	RC	FC
250	0.8170	**0.8419**	**0.8954**
500	0.7900	0.7252	0.8832
750	**0.8609**	0.4554	0.8046
1000	0.8378	0.3341	0.8108

Apêndice C.

Abaixo, encontram-se listadas as 40 perguntas que foram empregadas para a avaliação neste estudo. Essas perguntas foram cuidadosamente elaboradas pelos autores com base nos contextos extraídos, assegurando que cada uma delas fosse respondível com as informações disponíveis. Apenas foram selecionadas perguntas para as quais havia algum contexto que pudesse fornecer uma resposta adequada, garantindo assim a relevância e a precisão das avaliações realizadas.

Para garantir a qualidade e a relevância das perguntas, os autores também consultaram FAQs sobre oftalmologia, buscando inspiração nas dúvidas frequentes relatadas por pacientes. Este processo de pesquisa ajudou a garantir que as perguntas refletissem questões reais e pertinentes, baseadas em situações comuns encontradas na prática oftalmológica.

É importante notar que os contextos utilizados para responder as perguntas não podem ser disponibilizados publicamente, pois foram extraídos exclusivamente para fins acadêmicos. Esta restrição visa proteger a integridade e a confidencialidade das informações, assegurando que os dados sejam utilizados apenas para os propósitos da pesquisa e não sejam divulgados fora do âmbito acadêmico.

TABLE 12. TABELA CONTENDO TODAS AS PERGUNTAS AVALIADAS NO ESTUDO

Pergunta
Olá! Tenho diabetes e li que isso pode afetar minha visão. Quais são os cuidados especiais que devo tomar para proteger meus olhos?
Bom dia! Eu trabalho em frente a um computador o dia todo e tenho experimentado visão turva e dores de cabeça frequentes. Existe algum programa de exercícios para os olhos que possa ajudar?
Oi! Minha mãe está ficando mais velha e tem dificuldade em ler textos pequenos. Existe algum tipo de dispositivo de ampliação que você recomendaria para facilitar a leitura?
Boa tarde! Tenho alergia sazonal e meus olhos ficam extremamente coçando e vermelhos durante essa época do ano. Existe algum colírio específico que possa ajudar a aliviar esses sintomas?
Olá! Tenho um filho pequeno que usa óculos, mas ele está resistindo a usá-los na escola. Existe alguma estratégia que eu possa usar para incentivá-lo a usar os óculos?
Bom dia! Eu pratico esportes regularmente e estou preocupado com o risco de lesões oculares. Existe algum tipo de protetor ocular que você recomendaria para atividades esportivas específicas?
Olá, eu tenho notado que minha visão está ficando embaçada recentemente. Quais podem ser as possíveis causas desse problema?
Bom dia! Eu uso óculos há anos, mas ultimamente tenho sentido dores de cabeça frequentes ao usá-los. Isso poderia significar que preciso de uma nova prescrição?
Boa tarde! Minha avó foi diagnosticada com catarata. Gostaria de entender melhor sobre o procedimento cirúrgico e como posso ajudá-la durante a recuperação.
Oi! Minha filha está tendo dificuldade para se concentrar na escola e diz que às vezes vê manchas escuras flutuando em sua visão. Isso é algo para se preocupar?
Boa noite! Eu trabalho muitas horas seguidas no computador e ultimamente tenho sentido os olhos secos e irritados. Existem algumas medidas que eu possa tomar para aliviar esses sintomas?
Olá, tenho notado que meus olhos ficam vermelhos com frequência, especialmente após passar muito tempo ao ar livre. Existe alguma preocupação com isso?
Bom dia! Tenho uma viagem planejada para uma região muito ensolarada e estou preocupado com a exposição excessiva ao sol. Além dos óculos de sol, o que mais posso fazer para proteger meus olhos?
Oi! Eu tenho um filho pequeno que está constantemente esfregando os olhos e piscando com frequência. Isso pode indicar algum problema ocular?
Olá, recentemente comecei a ver halos em torno das luzes à noite. Isso poderia indicar algum problema ocular?
Bom dia! Tenho um histórico familiar de glaucoma e estou preocupado com o risco de desenvolvê-lo. Existem medidas preventivas que posso tomar para reduzir minhas chances?
Oi! Eu costumo usar maquiagem nos olhos diariamente, mas tenho notado que eles ficam irritados com mais frequência. Existe algum tipo específico de maquiagem que seja mais seguro para os olhos?
Boa tarde! Tenho astigmatismo e estou pensando em experimentar lentes de contato. Quais são os prós e contras em comparação com os óculos?
Olá! Eu trabalho em um ambiente com ar condicionado o dia todo e tenho notado que meus olhos ficam vermelhos e secos com mais frequência. Existe alguma maneira de evitar isso?
Posso tomar banho de rio ou de piscina com lente de contato?
Posso tomar banho de piscina de olho aberto?
Não usar óculos, faz o grau aumentar?
Sempre tive a visão normal, por que agora preciso esticar o braço para ver para perto?
Com que frequência se deve fazer exame médico oftalmológico?
Não utilizar os óculos e ”forçar a vista” piora o ”grau dos olhos”?
Posso lavar os olhos com água boricada ou soro fisiológico?
Visão dupla é problema oftalmológico ou neurológico?
O que causa o escurecimento da visão quando se levanta rapidamente?
A partir de que idade a vista começa a ”ficar cansada”?
Como devo limpar os olhos do bebê?
Existem contraindicações para o uso de lentes de contato?
A cirurgia de catarata pode ser feita com laser?
O que é Cirurgia Refrativa?
O que é ceratocone?
Como escolher os óculos de sol adequados para proteger os olhos?
Quais são os principais fatores de risco para degeneração macular relacionada à idade?
Quais são os benefícios e riscos da cirurgia de correção de miopia?
Como posso cuidar adequadamente das lentes de contato?
Quais são os efeitos do diabetes na visão e como posso proteger meus olhos se tiver diabetes?

Referencias

[1] L. Assi, F. Chamseddine, P. Ibrahim, H. Sabbagh, L. Rosman, N. Congdon, J. Evans, J. Ramke, H. Kuper, M. J. Burton, J. R. Ehrlich, and B. K. Swenor, “A Global Assessment of Eye Health and Quality of Life: A Systematic Review of Systematic Reviews,” *JAMA Ophthalmology*, vol. 139, pp. 526–541, 05 2021.

[2] W. H. Organization, *World report on vision*. World Health Organization, 2019.

[3] F. Antaki, S. Touma, D. Milad, J. El-Khoury, and R. Duval, "Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings," *Ophthalmology Science*, vol. 3, no. 4, p. 100324, 2023.

[4] I. A. Bernstein, Y. V. Zhang, D. Govil, I. Majid, R. T. Chang, Y. Sun, A. Shue, J. C. Chou, E. Schehlein, K. L. Christopher, S. L. Groth, C. Ludwig, and S. Y. Wang, "Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions," *JAMA Network Open*, vol. 6, pp. e2330320–e2330320, 08 2023.

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

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

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

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

[9] M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan, "Llm-based nlg evaluation: Current status and challenges," 2024.

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

[11] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[12] E. Sulem, O. Abend, and A. Rappoport, "BLEU is not suitable for the evaluation of text simplification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds.), (Brussels, Belgium), pp. 738–744, Association for Computational Linguistics, Oct.-Nov. 2018.

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

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

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

[16] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," 2024.

[17] X. Zhang, X. Ma, P. Shi, and J. Lin, "Mr. TyDi: A multi-lingual benchmark for dense retrieval," in *Proceedings of the 1st Workshop on Multilingual Representation Learning* (D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, and G. G. Sahin, eds.), (Punta Cana, Dominican Republic), pp. 127–137, Association for Computational Linguistics, Nov. 2021.

[18] P. Zweigenbaum, S. Sharoff, and R. Rapp, "Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora," in *Proceedings of the 10th Workshop on Building and Using Comparable Corpora* (S. Sharoff, P. Zweigenbaum, and R. Rapp, eds.), (Vancouver, Canada), pp. 60–67, Association for Computational Linguistics, Aug. 2017.

[19] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the Middle: How Language Models Use Long Contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 02 2024.

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

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

[22] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 27263–27277, Curran Associates, Inc., 2021.

[23] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. Huang, "A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets," in *Findings of the Association for Computational Linguistics: ACL 2023* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 431–469, Association for Computational Linguistics, July 2023.

[24] B. Bi, C. Li, C. Wu, M. Yan, W. Wang, S. Huang, F. Huang, and L. Si, "Palm: Pre-training an autoencodingautoregressive language model for context-conditioned generation," 2020.

[25] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," in *International Conference on Learning Representations*, 2019.

[26] M. Dunn, L. Sagun, M. Higgins, V. U. Guney, V. Cirik, and K. Cho, "Searchqa: A new qa dataset augmented with context from a search engine," 2017.

[27] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (K. Knight, A. Nenkova, and O. Rambow, eds.), (San Diego, California), pp. 110–119, Association for Computational Linguistics, June 2016.

[28] L. Gao and J. Callan, "Condenser: a pre-training architecture for dense retrieval," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Online and Punta Cana, Dominican Republic), pp. 981–993, Association for Computational Linguistics, Nov. 2021.

[29] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 6769–6781, Association for Computational Linguistics, Nov. 2020.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[32] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-shot learning with retrieval augmented language models," 2022.