INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PICOS
CURSO ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

EMANUELLE DE CARVALHO BRITO

GERAÇÃO AUTOMÁTICA DE LAUDOS DE GLAUCOMA POR LLMS: ESTUDO EXPERIMENTAL BASEADO EM EXAMES DE CAMPIMETRIA COMPUTADORIZADA

PICOS
2026

EMANUELLE DE CARVALHO BRITO

GERAÇÃO AUTOMÁTICA DE LAUDOS DE GLAUCOMA POR LLMS: ESTUDO EXPERIMENTAL BASEADO EM EXAMES DE CAMPIMETRIA COMPUTADORIZADA

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

Orientador: Prof. Dr. Rogério Figueredo de Sousa
Coorientador: Prof. Me. Anthony Irlan Marques Luz

PICOS
2026

Agradecimentos

Em primeiro lugar, agradeço ao meu maravilhoso Deus pelo privilégio da vida, pela força, amor e cuidado ao me conduzir até aqui. Sem Ele nada disso seria possível e nada teria conseguido. Nos dias difíceis e tempestuosos, Sua presença foi, é e sempre será o melhor refúgio para mim.

Agradeço à minha linda família pelo amor, apoio e lugar de conforto, onde posso me ancorar quando tudo parece estar rodando. Meus pais, irmãs, cunhados, sobrinhas, tias e primas. Sou grata a cada uma dessas pessoas maravilhosas que me fizeram ser quem sou hoje e me ajudaram a conseguir passar por toda a jornada até aqui.

Agradeço aos meus amigos muito especiais que tornaram a caminhada muito mais leve e divertida. A amizade, as conversas, risadas e brincadeiras foram a base para a perseverança de não desistir e lutar pelos sonhos e objetivos.

Aos meus professores, que são fonte de conhecimento e sabedoria. Ao meu orientador, pela pessoa maravilhosa que é e pelo excelente professor que desde as primeiras aulas me fez admirá-lo muito pela sua inteligência e carisma. Ao meu coorientador, por ter me ajudado muito no desenvolvimento deste trabalho e pela pessoa carismática, inteligente, divertida e cheia de ideias que é.

A todos que fizeram parte desta trajetória, este trabalho é uma conquista conjunta. Meu profundo e sincero agradecimento.

GERAÇÃO AUTOMÁTICA DE LAUDOS DE GLAUCOMA POR LLMS: ESTUDO EXPERIMENTAL BASEADO EM EXAMES DE CAMPIMETRIA COMPUTADORIZADA

Emanuelle de Carvalho Brito[1]
Rogério Figueredo de Sousa[2]
Anthony Irlan Marques Luz[3]

RESUMO

De acordo com a Organização Mundial de Saúde (OMS), o glaucoma é uma das principais causas de cegueira irreversível no mundo, sendo o diagnóstico precoce essencial para reduzir a progressão da doença. Nesse contexto, a geração automática de laudos médicos por meio de modelos de linguagem de grande porte (*Large Language Models – LLMs*) apresenta potencial para auxiliar especialistas e otimizar o fluxo de trabalho clínico. Este trabalho propõe a utilização de LLMs para a geração automática de laudos de glaucoma a partir de exames de campimetria computadorizada. Para isso, foram avaliados os modelos Llama-3.3-70B-Versatile, GPT-OSS-20B e Gemini-2.5-Flash, utilizando um conjunto de 118 exames representados em formato JSON. Os laudos gerados foram comparados aos laudos de referência por meio das métricas BLEU, ROUGE-1, ROUGE-L, Similaridade Semântica e da abordagem *LLM-as-a-Judge.* Os resultados indicaram que o modelo Llama-3.3-70B-Versatile apresentou melhor desempenho em relação ao tempo médio de geração (10,61s) por laudo e à similaridade semântica, enquanto o Gemini-2.5-Flash destacou-se na abordagem qualitativa realizada pelo LLM juiz. Além disso, a combinação de métricas automáticas e avaliação qualitativa permitiu uma análise abrangente da qualidade dos laudos produzidos, demonstrando o potencial dos LLMs como ferramenta de apoio à elaboração de laudos oftalmológicos.

Palavras-chave: Glaucoma. Campimetria Computadorizada. LLMs. Inteligência Artificial. Processamento de Linguagem Natural.

ABSTRACT

According to the World Health Organization (WHO), glaucoma is one of the leading causes of irreversible blindness worldwide, making early diagnosis essential to reduce disease progression. In this context, the automatic generation of medical reports using Large Language Models (LLMs) has the potential to support specialists and optimize clinical workflows. This study proposes the use of LLMs for the automatic generation of glaucoma reports based on computerized visual field examinations. Three language models—Llama-3.3-70B-Versatile, GPT-OSS-20B,

[1] Graduanda em Análise e Desenvolvimento de Sistemas. IFPI. emanuellee7britoo@gmail.com
[2] Professor Doutor em Processamento de Linguagem Natural. IFPI.
[3] Professor Mestre em Ciência da Computação. IFPI.

and Gemini-2.5-Flash—were evaluated using a dataset of 118 examinations represented in JSON format. The generated reports were compared with reference reports using the BLEU, ROUGE-1, ROUGE-L, Semantic Similarity metrics, and the LLM-as-a-Judge approach. The results showed that the Llama-3.3-70B-Versatile model achieved the best overall performance, presenting the lowest average report generation time and the highest semantic similarity to the reference reports. Furthermore, the combination of automatic evaluation metrics and qualitative assessment provided a comprehensive analysis of the generated reports, demonstrating the potential of LLMs as a support tool for ophthalmological report generation.

Keywords: Glaucoma. Computerized Perimetry. Large Language Models. Artificial Intelligence. Natural Language Processing.

Data de aprovação: 12/08/2026

1 INTRODUÇÃO

O glaucoma representa um grande desafio da oftalmologia, sendo a principal causa de cegueira irreversível no mundo. A meta-análise[4] realizada por Tham et al. (2014) estima que até o ano 2040 o número de indivíduos acometidos com a doença ultrapasse 100 milhões. No Brasil, os números variam entre 1,5 a 2 milhões de casos, com tendência de crescimento devido ao envelhecimento populacional (Guedes et al., 2025).

Caracterizado como uma neuropatia óptica progressiva, o glaucoma causa a degradação do nervo óptico, estrutura responsável por processar os estímulos visuais e construir imagens coerentes (Cardoso et al., 2025). Entre suas características clínicas, destacam-se a escavação do disco óptico e a perda do campo visual periférico, frequentemente associadas à elevação da pressão intraocular (PIO), podendo evoluir para perda total ou parcial da visão (Freitas et al., 2025).

A percepção do glaucoma pode ser realizada por meio da campimetria computadorizada, exame que avalia a sensibilidade do campo visual central e periférico a partir da resposta do paciente aos estímulos luminosos enviados pelo campímetro. O exame gera um mapa detalhado do campo visual, composto por representações numéricas, índices globais e mapas em escala de cinza, possibilitando a identificação de falhas e a análise da progressão do comprometimento visual do paciente (Baldon et al., 2025). Entretanto, devido a natureza assintomática da doença nos estágios iniciais, o diagnóstico pode apresentar variações conforme a interpretação clínica, dificultando a detecção precoce (Takahashi, et al., 2022).

Com o crescimento acelerado dos casos de glaucoma, a Inteligência Artificial (IA) tem se mostrado uma ferramenta importante no apoio ao diagnóstico da doença. As técnicas de IA permitem que os algoritmos aprendam padrões complexos a partir dos dados, possibilitando a análise eficiente e automatizada de exames médicos (Silva, 2022). Entre as abordagens de IA destaca-se o *Machine Learning* (ML), que possibilita a identificação automática de padrões a partir de um conjunto de dados, permitindo que os algoritmos aprendam e realizem previsões para tarefas

[4] Método estatístico usado para juntar os resultados de vários estudos científicos sobre o mesmo tema.

específicas (Paixão et al., 2022).

Dentro do campo de ML está o *Deep Learning* (DL), que utiliza redes neurais artificiais com múltiplas camadas para aprender padrões e características complexas (Barbosa;Portes, 2023). Russo et al. (2024) destacam que os algoritmos de DL superam oftalmologistas na detecção do glaucoma a partir da análise da campimetria computadorizada. Diante disso, este trabalho tem como objetivo avaliar comparativamente o desempenho de LLMs na geração de minutas de laudos a partir de dados estruturados extraídos de exames de campimetria computadorizada. A avaliação considera métricas de sobreposição lexical, similaridade semântica, avaliação automatizada por um LLM juiz e tempo de geração.

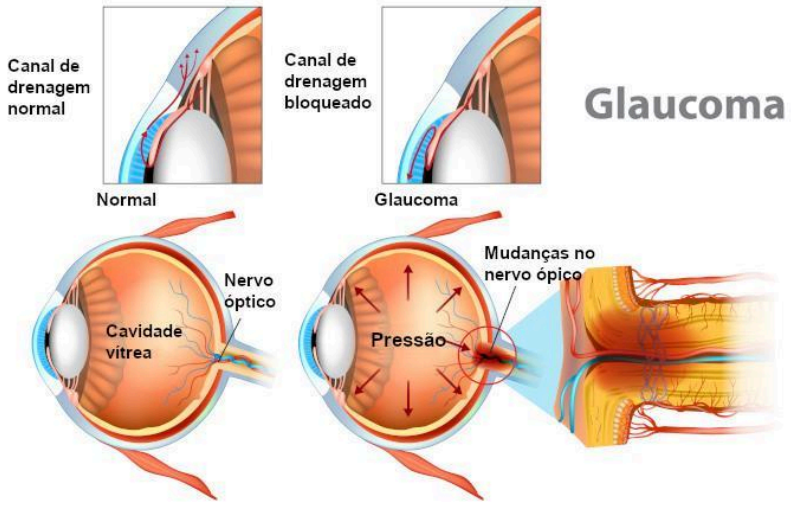
2 REFERENCIAL TEÓRICO

Esta seção apresenta os fundamentos teóricos que embasaram o desenvolvimento deste trabalho, contemplando o conceito de glaucoma, campimetria computadorizada como método de diagnóstico e LLMs no glaucoma. Além disso, são discutidos estudos relacionados à aplicação de LLMs na análise de exames de campimetria computadorizada.

2.1 CONCEITO DE GLAUCOMA

O glaucoma, por definição, é uma doença ocular crônica que danifica o nervo óptico, comprometendo o campo visual, principalmente o periférico, e pode causar a perda irreversível da visão (Silva, 2024). Geralmente é associado ao aumento da PIO e apresenta sintomas comuns entre os pacientes, como a escavação do nervo óptico e a perda do campo visual periférico (Curado et al., 2023). A Figura 1 mostra a distinção entre um olho glaucomatoso e um não glaucomatoso. Em condições normais, o humor aquoso é drenado adequadamente, mantendo a PIO estável. No glaucoma, o canal de drenagem é bloqueado, impedindo o fluxo do humor aquoso, que se acumula na cavidade, ocasionando a elevação da PIO.

Figura 1 - Formação do glaucoma



Fonte: Santos e Sardinha (S.d)

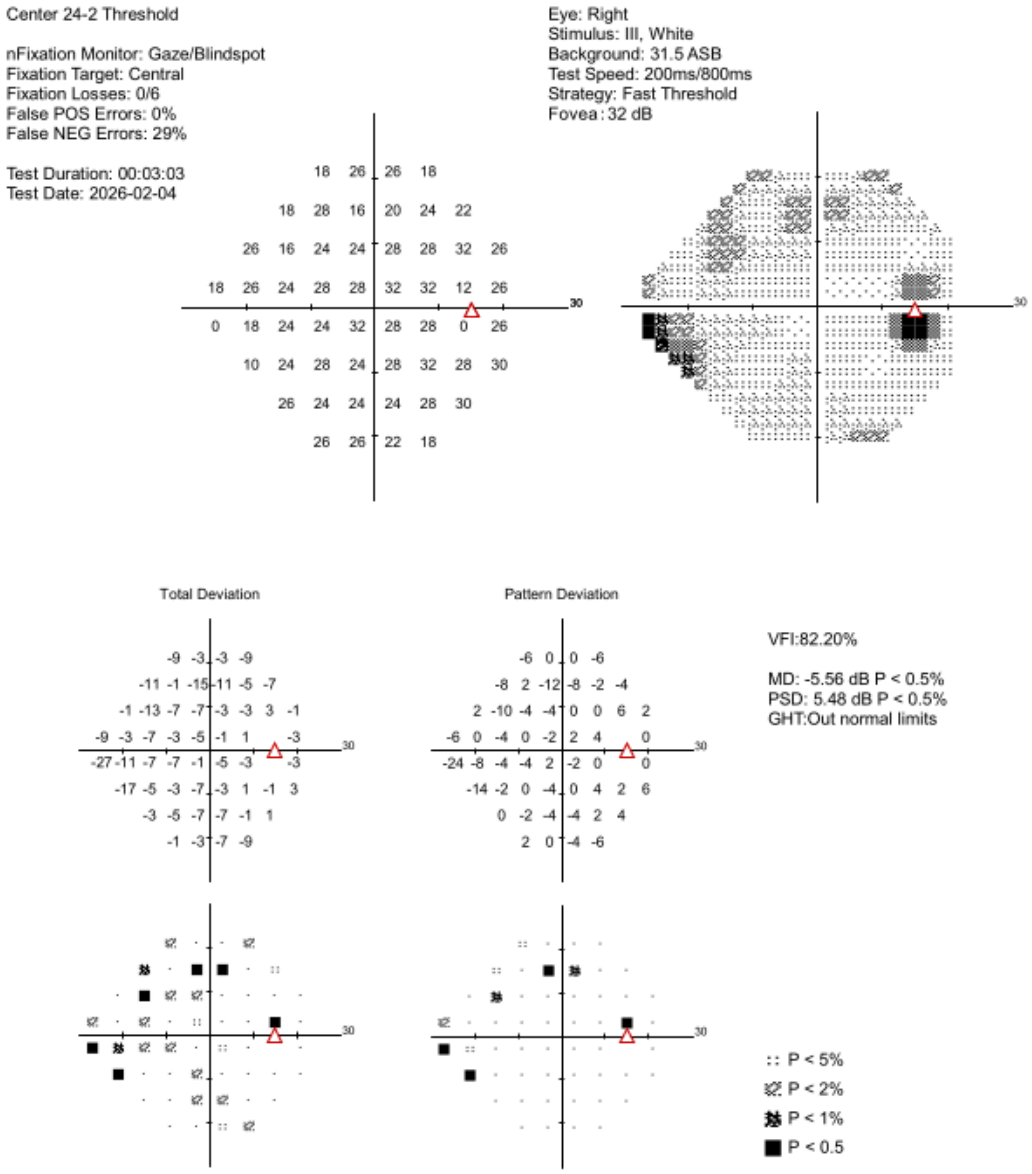
Como ilustrado na Figura 1, as alterações no sistema de drenagem do humor aquoso podem provocar a elevação da PIO, contribuindo para a progressão gradual da doença. A degeneração do nervo óptico ocorre de maneira gradativa e, na maioria dos casos, assintomática nos estágios iniciais, fatores que contribuem para o diagnóstico tardio e perda irreversível da visão (Gomes et al., 2025).

2.2 CAMPIMETRIA COMPUTADORIZADA COMO MÉTODO DE DIAGNÓSTICO

Uma das estratégias para a detecção do glaucoma é a campimetria computadorizada, exame que avalia a sensibilidade do campo visual central e periférico a partir da resposta do paciente aos estímulos luminosos nas regiões do campo visual (Chaoubah, 2017). O exame gera um mapa detalhado do campo visual, composto por representações numéricas, índices globais e mapas em escala de cinza, possibilitando a identificação de falhas e a análise da progressão do comprometimento visual do paciente (Baldon et al., 2025).

A campimetria é fundamental para a avaliação do campo visual do paciente, permitindo o mapeamento computadorizado das áreas de sensibilidade através da resposta do paciente aos estímulos enviados (Cemo, s.d). A Figura 2 apresenta o resultado do exame através de gráficos em escala de cinzas do desvio padrão e desvio total, desvio médio (MD) e desvio padrão do padrão (PSD), além dos índices de confiabilidade como falsos positivos e falsos negativos (Borges et al., 2005).

Figura 2 - Campo visual



Fonte: exame anonimizado disponibilizado pela clínica parceira

O diagnóstico se dá através da análise dos índices e gráficos apresentados no campo visual. Silva (2013) destaca os principais índices do campo visual:

- **MD (*Mean Deviation*):** sensibilidade geral do campo visual. Representa a média ponderada dos valores do gráfico de desvio total. O MD pode ser positivo, indicando sensibilidade acima da média.
- **PSD (*Pattern Standard Deviation*):** desvio padrão da média dos valores do gráfico de desvio padrão (*Pattern Deviation*). É um indicador de defeitos

localizados no campo visual.

- **GHT (*Glaucoma Hemifield Test*):** teste que compara cinco zonas inferiores com cinco zonas superiores do campo com o objetivo de detectar assimetrias.

Conforme explica o Ministério da Saúde (Brasil, 2023), o principal objetivo da campimetria computadorizada é identificar, por meio da presença de escotomas (ponto cego ou uma área de visão reduzida) e aumento da mancha cega, a perda ou comprometimento do campo visual central e periférico. Russo, et al. (2024) afirmam que a campimetria computadorizada é o método mais preciso para avaliar o campo visual do paciente no glaucoma.

2.3 LLMS NO GLAUCOMA

A inteligência artificial se mostrou relevante na área da saúde, demonstrando precisão desde a identificação até o diagnóstico de doenças. Segundo Dunningham e Filho (2023), a IA se tornou notável no auxílio ao diagnóstico de doenças pela sua capacidade de coletar, analisar e processar grandes volumes de dados, identificando padrões e informações relevantes em exames médicos. Uma de suas classificações é IA generativa que, baseada em modelos de aprendizado profundo, analisa, processa os dados e interage com os usuários, sendo muito utilizada na criação de textos e imagens (Mattos;Valente, 2026).

No âmbito de IA generativa, os LLMs (*Large Language Models)* são modelos de linguagem treinados com grandes volumes de dados para prever sequências de palavras ou sentenças, dado um contexto anterior (Figênio;Gomes-Jr, 2023). Esses modelos representam um suporte tecnológico para o ser humano, visto que conseguem gerar textos, fazer análises, traduções e demais tarefas a partir de instruções dadas pelo usuário (Ribeiro;Tsunoda, 2025). Na oftalmologia, os modelos de linguagem são utilizados para documentação clínica, orientação e resposta a perguntas relacionadas a doenças (Rubegni et al., 2026).

Entre os estudos que investigam a aplicação de modelos de linguagem na oftalmologia, destaca-se o trabalho de Tan et al. (2024), que avaliou o desempenho do *ChatGPT* (GPT-3.5) na resposta a perguntas frequentes de pacientes sobre glaucoma. Os autores elaboraram 24 questões clínicas distribuídas entre diagnóstico, tratamento, cirurgias e emergências oftalmológicas, cujas respostas

foram avaliadas por três especialistas em glaucoma quanto à precisão, abrangência e segurança. Os resultados mostraram que 70,8% das respostas foram consideradas adequadas, principalmente relacionadas ao diagnóstico, e que o desempenho do modelo melhorou significativamente após um processo de autocorreção, evidenciando o potencial dos LLMs como ferramenta de apoio à orientação de pacientes.

Xue et al. (2024) propuseram o *Xiaoqing*, um modelo de perguntas e respostas sobre glaucoma baseado em modelos de linguagem de grande porte (LLMs). O desempenho do modelo foi avaliado por meio da comparação de suas respostas a 22 perguntas com as geradas por quatro modelos existentes (*ChatGPT-3*.5, *ChatGPT-4*, *Huatuo* e *Ivy*). Os resultados demonstraram que o *Xiaoqing* apresentou melhor desempenho em termos de informatividade e legibilidade, além de fornecer respostas mais específicas e adequadas ao contexto do glaucoma em comparação aos modelos de propósito geral.

Os estudos apresentados evidenciam o potencial dos modelos de linguagem na oftalmologia, especialmente em tarefas relacionadas ao glaucoma, como a resposta a perguntas e o suporte à orientação de pacientes. Embora apresentados resultados promissores, observa-se que as pesquisas concentram-se principalmente em sistemas de perguntas e respostas, sendo ainda limitadas as investigações sobre a utilização desses modelos em outras tarefas clínicas, como a interpretação de exames e a geração de documentos médicos. Nesse contexto, este trabalho busca contribuir para essa lacuna ao investigar a aplicação de modelos de linguagem na geração automática de laudos de glaucoma a partir de exames de campimetria computadorizada ao realizar uma avaliação comparativa do desempenho de diferentes LLMs por meio de métricas de Processamento de Linguagem Natural e da abordagem *LLM-as-a-Judge.*

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa foi desenvolvida por meio de uma abordagem experimental, visando avaliar a utilização de LLMs para geração automática de laudos a partir de exames de campimetria computadorizada. Para isso, foram realizadas as etapas de coleta e pré-processamento de dados, implementação dos modelos, geração dos laudos e avaliação dos resultados obtidos através de métricas de processamento de linguagem natural e LLM como juiz.

3.1 COLETA E PRÉ-PROCESSAMENTO DE DADOS

Inicialmente, os dados foram obtidos através de um *dataset* privado disponibilizado por uma clínica oftalmológica parceira, composto por 118 exames de campimetria computadorizada. Os exames foram disponibilizados no formato pdf, cuja maioria apresenta duas imagens de exame (cada imagem correspondente a um olho) e uma imagem de laudo gerado pelo aparelho.

Após a coleta, foram utilizadas funções específicas para a extração de informações textuais dos exames e laudos, garantindo a anonimização ao extrair apenas os campos necessários para análise. Os dados textuais dos exames foram armazenados em arquivos em formato JSON, com campos previamente definidos, enquanto as informações textuais dos laudos foram armazenadas em arquivos de texto.

3.2 IMPLEMENTAÇÃO DOS MODELOS E GERAÇÃO DE LAUDOS

Para a etapa de geração de laudos, foram escolhidos três modelos de linguagem: ***Llama-3.3-70B-Versatile*** e ***GPT-OSS-20B***, ambos disponibilizados pela API do Groq, e ***Gemini-2.5-Flash***, disponibilizado pela API do Google. A escolha foi motivada pela capacidade de processamento de linguagem natural e disponibilidade para utilização pela API, visto que modelos locais demandam alto custo computacional. Além disso, os modelos utilizados contam com bilhões de parâmetros, o que amplia a capacidade e qualidade ao analisar os exames e gerar os laudos.

Cada modelo recebeu 118 exames de campimetria computadorizada,

organizados em um arquivo JSON. Os arquivos continham as informações textuais extraídas dos exames, como índices de confiabilidade e índices globais de danos. Além dos dados do exame, foi fornecido um *prompt* com instruções específicas sobre o comportamento esperado do modelo, incluindo a linguagem técnica compatível com a oftalmologia, estrutura padronizada dos laudos, objetividade na descrição dos achados e restrição à geração de informações não contidas nos exames, com o objetivo de reduzir as alucinações.

Durante a execução, os modelos percorreram cada arquivo em formato JSON do diretório de entrada, processando individualmente cada arquivo e gerando o laudo correspondente ao exame em formato TXT. Os laudos gerados foram armazenados em diretórios específicos para cada modelo, possibilitando a avaliação posterior.

3.3 AVALIAÇÃO DE DESEMPENHO

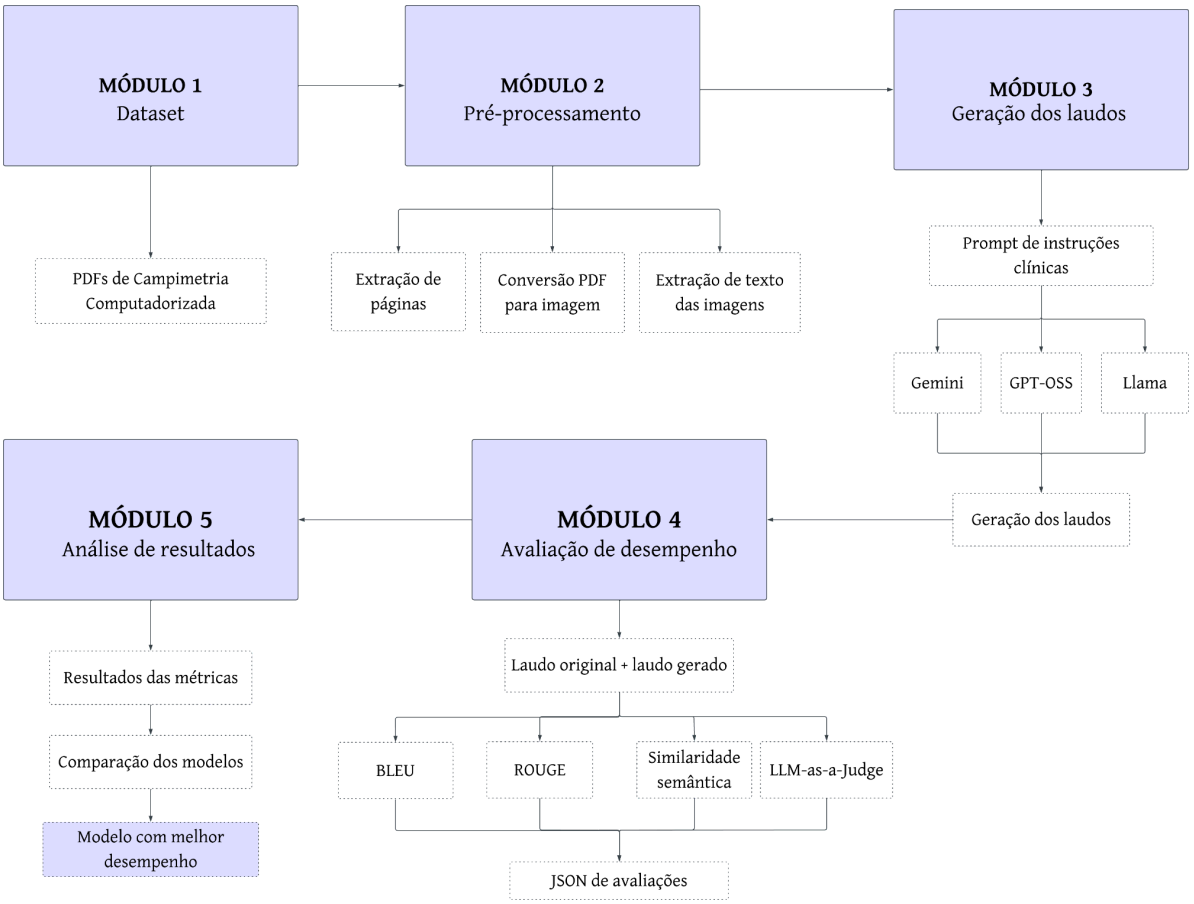
Após a geração dos laudos, foi realizada a avaliação dos modelos por meio da comparação entre os laudos gerados e os laudos originais. Para isso, foram utilizadas métricas de PLN, com o objetivo de mensurar a similaridade textual e semântica, além da avaliação qualitativa pela abordagem ***LLM-as-a-Judge***. Com base na conceituação de Faé (2025) a respeito da avaliação de modelos de linguagem, foram utilizadas as seguintes métricas:

- **BLEU (*Bilingual Evaluation Understudy*):** mede a similaridade entre o laudo gerado e o laudo de referência por meio da correspondência entre sequências de palavras (*n-grams*). Quanto maior a pontuação, maior a sobreposição textual entre os documentos.
- **ROUGE-1 (*Recall-Oriented Understudy for Gisting Evaluation*):** avalia a sobreposição de unigramas entre o texto gerado e o texto de referência, sendo amplamente utilizada para medir a cobertura das informações presentes no documento original.
- **ROUGE-L:** baseia-se na maior subsequência comum (*Longest Common Subsequence – LCS*) entre os textos, permitindo avaliar a preservação da estrutura e da sequência das informações.
- **Similaridade semântica:** mede o grau de proximidade de significado entre o

laudo gerado e o laudo de referência por meio de representações vetoriais (*embeddings*), permitindo identificar semelhanças mesmo quando são utilizadas palavras ou construções textuais diferentes.

- ***LLM-as-a-Judge*:** consiste na utilização de um modelo de linguagem para avaliar qualitativamente os laudos gerados. O modelo analisa critérios como precisão clínica, completude, coerência, clareza, conformidade com as informações presentes no exame e ocorrência de alucinações, atribuindo uma avaliação ao laudo produzido.

Figura 3 - Pipeline de execução



Fonte: autoria própria

A Figura 3 apresenta a estrutura do código, composta pelas etapas metodológicas descritas anteriormente. A combinação de métricas automáticas e da abordagem *LLM-as-a-Judge* permitiu avaliar tanto a similaridade textual entre os laudos quanto sua qualidade clínica, permitindo uma análise mais abrangente do desempenho dos modelos.

4 RESULTADOS E DISCUSSÕES

Nesta seção são apresentados e discutidos os resultados obtidos na avaliação dos modelos de linguagem utilizados para a geração de laudos de glaucoma. O desempenho dos modelos foi analisado por meio das métricas BLEU, ROUGE-1, ROUGE-L, Similaridade semântica e *LLM-as-a-Judge*. Além das métricas de PLN, foi analisado o tempo médio em segundos de cada modelo na geração dos laudos, conforme mostrado na Tabela 1.

Tabela 1 - Comparação de desempenho dos modelos utilizados na geração dos laudos

Modelo	Qtd. laudos gerados	Média de tempo por laudo (s)	BLEU	ROUGE-1	ROUGE-L	Similaridade semântica	Confiabilidade por LLM juiz
Llama-3.3-70B-Versatile	118	10,61	0,01	0,40	0,20	0,67	80,2%
GPT-OSS-20B	118	20,58	0,02	0,42	0,23	0,65	75,2%
Gemini-2.5-Flash	118	22,22	0,01	0,41	0,24	0,66	86,3%

Fonte: autoria própria

Conforme apresentado na Tabela 1, o modelo ***Llama-3.3-70B-Versatile***, disponibilizado pela *API* do *Groq*, alcançou melhores resultados em relação ao tempo médio de geração por laudo e à similaridade semântica. O modelo gerou os 118 laudos em um tempo médio de 10.61 segundos por exame, aproximadamente metade do tempo observado nos demais modelos, evidenciando maior eficiência computacional. Em relação à similaridade semântica, o modelo apresentou pontuação ligeiramente maior, embora os valores foram consideravelmente baixos. Os modelos receberam os dados estruturados dos exames sem exemplos prévios de laudos no prompt, o que contribuiu para os valores baixos de similaridade semântica.

Os modelos apresentaram valores reduzidos de BLEU, indicando baixa correspondência lexical entre os laudos. Isso significa que poucas palavras ou sentenças dos laudos gerados foram sobrepostas dos laudos originais. É importante ressaltar que valores baixos de BLEU não significam laudos ruins, mas sim que, em

um mesmo contexto foram usadas palavras distintas, a exemplo “glaucoma” e “dano glaucomatoso”. Ambos indicam conteúdo semelhante com palavras diferentes. A interpretação conjunta de similaridade semântica e BLEU indica que o conteúdo principal do laudo foi preservado, embora as sentenças não se sobrepuseram.

A métrica ROUGE-1 apresentou valores aproximados e relativamente baixos, entre 0.40 e 0.42. Isso indica que, em média, existe aproximadamente 42,8% de sobreposição de palavras entre o texto original e o de referência. Em relação à ROUGE-L, os valores entre 0.20 e 0.24 representam o equilíbrio (F1-Score) entre a precisão e a revocação da maior subsequência de palavras entre o texto gerado e o de referência. Ambos os comportamentos são esperados de LLMs, visto que o conteúdo é preservado, enquanto as palavras e sentenças são alteradas conforme o vocabulário do modelo.

Por fim, a avaliação qualitativa foi realizada por meio da abordagem ***LLM-as-a-Judge***, representada pela coluna **Confiabilidade por LLM Juiz**. Nessa etapa, o modelo ***GPT-OSS-120B*** comparou individualmente os laudos gerados com os respectivos laudos de referência. Para isso, foi elaborado um prompt contendo critérios específicos de avaliação, incluindo a **fidelidade clínica**, responsável por verificar se o modelo preservou as informações do exame sem introduzir dados inexistentes, e a **confiabilidade médica**, destinada a avaliar a adequação da linguagem técnica, o uso correto da terminologia médica e a coerência clínica do laudo. Para cada critério, o modelo atribuiu uma nota de 0 a 10. Os resultados foram armazenados em arquivos no formato JSON, organizados de acordo com o modelo avaliado e as respectivas pontuações obtidas. Posteriormente, foi calculada a média aritmética das notas atribuídas aos laudos de cada modelo. Em seguida, esse valor foi convertido para uma escala percentual (0 a 100%), originando o índice médio de confiabilidade apresentado na coluna **Confiabilidade por LLM Juiz**.

3 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma abordagem de geração de laudos de glaucoma a partir do exame de campimetria computadorizada utilizando LLMs. Para isso, foram avaliados os modelos: *Llama-3.3-70B-Versatile, GPT-OSS-20B* e *Gemini-2.5-Flash,* considerando métricas de Processamento de Linguagem Natural.

Os resultados demonstraram que, embora utilizados apenas 118 exames, os modelos foram capazes de gerar laudos condizentes com as informações presentes nos exames. Em relação às métricas quantitativas, o *Llama-3.3-70B-Versatile* destacou-se pela sua capacidade de análise e geração de laudos em menor tempo. Em contrapartida, a métrica LLM-as-a-Judge mostrou que o *Gemini-2.5-Flash* apresentou melhores resultados qualitativos e técnicos. Os valores baixos de algumas métricas apontam a capacidade dos modelos de preservarem o conteúdo, apesar da utilização de estruturas lexicais diferentes.

Além da comparação entre diferentes modelos de linguagem, este trabalho demonstrou a viabilidade da utilização desses modelos na análise dos exames de campimetria e geração de laudos correspondentes. A combinação de métricas qualitativas e quantitativas possibilitou uma avaliação abrangente da capacidade de análise dos exames e qualidade dos laudos gerados, contribuindo para pesquisas voltadas à aplicação da Inteligência Artificial na oftalmologia.

Como limitações, destaca-se a utilização de um conjunto de dados composto por 118 exames e avaliação restrita a três modelos de linguagem. Além disso, apesar da análise qualitativa pelo LLM juiz apresentar resultados promissores, estes devem ser monitorados por especialistas em oftalmologia, para garantir maior confiabilidade e precisão.

Como trabalhos futuros, pretende-se ampliar o conjunto de dados, avaliar modelos multimodais para análise de texto e imagens, fazer experimentos com modelos específicos para tarefas médicas, como o MedGemma, e aplicar média ponderada aos dados do exame, de acordo com níveis de prioridade, com o objetivo de melhorar a precisão dos laudos gerados. Além disso, técnicas de engenharia de prompt e ajuste fino são estratégias abordadas futuramente, visando aprimorar a qualidade e confiabilidade dos laudos.

REFERÊNCIAS

BALDON, L. V.; GLÓRIA, I. R. da; MARTUCHELI, M. de O.; MARTINS, A. D. F.; PEREIRA, D. A. GLAUCOMA: MECANISMOS DA DOENÇA E IMPORTÂNCIA DO DIAGNÓSTICO PRECOCE. REVISTA FOCO, *[S. l.]*, v. 18, n. 9, p. e9724, 2025.

BARBOSA, L. M.; PORTES, L. A. F. A inteligência artificial. Revista Tecnologia Educacional, Rio de Janeiro, n. 236, p. 16-27, 2023.

BORGES, Kleber Fragoso; INOCÊNCIO, Magda Passamani; BASTOS, Mônica de Araújo Pereira; INOCÊNCIO, Virgínia Camargo; TARDIN, João Roberto Garcia; SILVEIRA, Luiz Felipe Queiroz Sampaio da. Perimetria de dupla frequência Humphrey Matrix: avaliação do tempo de exame sem e com correção. Revista Brasileira de Oftalmologia, Rio de Janeiro, v. 64, n. 5, p. 319-324, 2005.

BRASIL. Ministério da Saúde. Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde (CONITEC). Protocolo clínico e diretrizes terapêuticas do glaucoma. Brasília, DF: Ministério da Saúde, 2023. Disponível em: https://www.gov.br/conitec/pt-br/midias/relatorios/2023/protocolo-clinico-e-diretrizes-terapeuticas-do-glaucoma.pdf. Acesso em: 21 fev. 2026.

CARDOSO, P. H. D. S. M.; SOUSA, L. R.; ALMEIDA SANTOS, T. F.; SILVA MONTEIRO, M.; HAUAJI, D. D. F. F.; SILVA VIEIRA, K. A.; et al. Abordagem geral do glaucoma. Revista Eletrônica Acervo Saúde, v. 25, n. 5, e20192, 2025.

CHAOUBAH, Alfredo. Desenvolvimento de equipamento para exame de campo visual. *HU Revista*, Juiz de Fora, v. 43, n. 3, p. 265–268, jul./set. 2017. DOI: 10.34019/1982-8047.2017.v43.2823.

CLÍNICA CEMO. Campimetria computadorizada. Clínica CEMO, [s.d.]. Disponível em: https://cemoosasco.com.br/exames/campimetria-computadorizada/. Acesso em: 17 jul. 2026.

CURADO, S. C. C.; PAIVA, I. B. Glaucoma: uma revisão bibliográfica. Research, Society and Development, v. 12, n. 12, e13121243731, 2023.

GOMES, C. D.; REIS KANAWA, A. A.; DE ALMEIDA RESENDE, J.; TAVARES TESTONI, N. M.; DE ARAÚJO GOMES DA SILVA, A.; DE OLIVEIRA REBOUÇAS, E.; MIGUEL CUADROS, S.; CARVALHO ENCINAS, B.; TOMASI, G.; SANT'ANA SOARES, H.; BONJOUR MENDES, A. L. Glaucoma de Ângulo Aberto: Abordagem Atual, Desafios Diagnósticos e Avanços Terapêuticos. Brazilian Journal of Implantology and Health Sciences, *[S. l.]*, v. 7, n. 8, p. 886–898, 2025.

DUNNINGHAM, William; ANDRADE FILHO, Antônio. Inteligência artificial (IA) na formação e na prática médica. *Revista Brasileira de Neurologia e Psiquiatria*, v. 27, n. 3, p. 2-3, set./dez. 2023.

FREITAS, G. dos S. G. F. de; MOREIRA, F. M.; MENDONÇA, A. C. M. Do glaucoma ao neuro glaucoma. Journal Archives of Health, *[S. l.]*, v. 6, n. 4, p. e3513, 2025.

FAÉ, Eduardo Dalmás. *Comparativo de métricas intrínsecas para avaliação do desempenho de modelos em open-ended tasks*. 2025. Monografia (Bacharelado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2025.

FIGÊNIO, Mateus R.; GOMES-JR, Luiz. Ética na era dos Modelos de Linguagem Massivos (LLMs): um estudo de caso do ChatGPT. *In*: ESCOLA REGIONAL DE BANCO DE DADOS (ERBD), 18. , 2023, Palmas/PR. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023 . p. 100-107.ISSN 2595-413X. DOI: https://doi.org/10.5753/erbd.2023.229510.

GUEDES, R. A. P. Terapias não farmacológicas do glaucoma de ângulo aberto: estamos vivendo uma mudança de paradigmas? Revista Brasileira de Oftalmologia, v. 84, e0001, 2025.

MATTOS, Alexandre Magalhães de; VALENTE, Tania Cristina de Oliveira. Inteligência artificial na área da saúde: desafios, possibilidades, regulamentação e implicações legais. *Revista do Direito Público*, Londrina, v. 21, n. 1, e49634, 2026.

PAIXÃO, G. M. D. M.; SANTOS, B. C.; ARAUJO, R. M. D.; RIBEIRO, M. H.; MORAES, J. L. D.; RIBEIRO, A. L. Machine learning na medicina: revisão e aplicabilidade. *Arquivos Brasileiros de Cardiologia*, v. 118, p. 95–102, 2022

RIBEIRO, Patrick Fernandes Rezende; TSUNODA, Denise Fukumi. Geração de textos dinâmicos e criativos com Grandes Modelos de Linguagem (LLMs): uma revisão sistemática. *Texto Livre*, Belo Horizonte, v. 18, e60103, 2025.

RUBEGNI, Giovanni; CARTOCCI, Alessandra; LUSCHI, Alessio; CASTELLINO, Niccolò; CAPPELLANI, Francesco; ROMANO, Dario; COLIZZI, Benedetta; ROSSETTI, Luca; TOSI, Gian Marco. Applications of Large Language Models in Glaucoma: A Scoping Review. *Vision*, v. 10, n. 1, art. 9, 2026.

RUSSO , L. M. M.; BRANDÃO , B. R. T. M.; FERREIRA, A. H. M.; LIMA , R. C. de; BIALLOWONS, L. L. S.; BIALLOWONS , S. S.; AURÉLIO , S. M.; EMERICK, F. T.; FREITAS, T. R. da S. de; FREITAS , G. P. de; GAMA , K. M. da; PINTO , D. N. dos S. ABORDAGENS INOVADORAS NO DIAGNÓSTICO PRECOCE DO GLAUCOMA. Brazilian Journal of Implantology and Health Sciences, *[S. l.]*, v. 6, n. 3, p. 1548–1562, 2024.

SANTOS, Vanessa Sardinha dos. Glaucoma: o que é, tipos, sintomas e tratamento. Mundo Educação, [s.d.]. Disponível em: https://mundoeducacao.uol.com.br/doencas/glaucoma.htm. Acesso em: 15 fev. 2026.

SILVA, Fabrício Reis da. Diagnóstico de glaucoma baseado em classificadores de aprendizagem de máquina utilizando dados do *Spectral Domain-OCT* e perimetria automatizada acromática. 2013. Dissertação (Mestrado) – Universidade Estadual de Campinas, Campinas, 2013.

SILVA, Fernanda Schäfer Tesch da. *Estudo da aplicação de deep learning como auxílio ao diagnóstico de glaucoma*. 2022. Trabalho de Conclusão de Curso

(Bacharelado em Engenharia Elétrica) — Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, RS, 2022.

SILVA, W. M. D. Análise de técnicas de explicabilidade em redes neurais convolucionais para diagnóstico de glaucoma, retinopatia diabética e catarata. 2024. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) — Universidade Federal de Campina Grande, Campina Grande, 2024.

TAKAHASHI, I. M.; PIRES, B. C.; REIS, L. S.; VIEIRA, J. M.; FERNANDES, C. R. Inteligência artificial no glaucoma: uma revisão literária. Revista Médica de Minas Gerais, v. 32, supl. 1, p. S31-S34, 2022.

TAN, Darren Ngiap Hao; THAM, Yih-Chung; KOH, Victor; SENG, Chee Loon; AQUINO, Maria Cecilia; LUN, Katherine; CHENG, Ching-Yu; NGIAM, Kee Yuan; TAN, Marcus. Evaluating Chatbot responses to patient questions in the field of glaucoma. *Frontiers in Medicine*, v. 11, art. 1359073, 8 July 2024.

THAM, Yih-Chung et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, v. 121, n. 11, p. 2081–2090, 2014. DOI: 10.1016/j.ophtha.2014.05.013.

XUE, Xiaojuan; ZHANG, Deshiwei; SUN, Chengyang; SHI, Yiqiao; WANG, Rongsheng; TAN, Tao; GAO, Peng; FAN, Sujie; ZHAI, Guangtao; HU, Menghan; WU, Yue. Xiaoqing: A Q&A model for glaucoma based on LLMs. *Computers in Biology and Medicine*, v. 174, 108399, May 2024.