Sintomas Linguísticos: Geração Aumentada por Recuperação e Raciocínio em LLMs sob a Variação Português–Inglês em Contextos Médicos

Guilherme Vianna de Moura, Gabriel Assis, Aline Paes

Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brazil

gvmoura@id.uff.br, assisgabriel@id.uff.br, alinepaes@ic.uff.br

Resumo

Modelos de Língua de Grande Porte (LLMs) têm demonstrado desempenho expressivo em tarefas de raciocínio médico. No entanto, sua robustez diante de variações linguísticas ainda é pouco explorada, especialmente em idiomas além do inglês, como o português. Neste trabalho, investigamos como o idioma de entrada afeta o desempenho e o comportamento de raciocínio de LLMs médicos, bem como se a Geração Aumentada por Recuperação (RAG) é capaz de mitigar eventuais limitações decorrentes dessas variações. Para isso, realizamos experimentos em português e em inglês, utilizando duas variantes do modelo MedGemma, com 4B e 27B parâmetros, e avaliando-as em três conjuntos de dados médicos. A avaliação combina métricas quantitativas de acurácia com análises qualitativas e estruturais das cadeias de raciocínio e das respostas geradas pelos modelos. Os resultados indicam que a variação linguística impacta de forma mais acentuada os modelos de menor porte. Em particular, a variante de 4B parâmetros apresenta desempenho consistentemente inferior quando as entradas são fornecidas em português. Em contraste, a variante de 27B parâmetros demonstra maior robustez entre idiomas, mantendo níveis semelhantes de acurácia e de estrutura de raciocínio tanto em português quanto em inglês. Embora o sistema de RAG implementado apresente recuperação de documentos de boa qualidade, sua integração não resulta em ganhos consistentes para o modelo menor, o que sugere limitações na exploração efetiva do contexto adicional. De forma geral, este trabalho contribui para o entendimento dos limites atuais dos LLMs médicos em contextos multilíngues, destacando os desafios associados ao desempenho em idiomas com recursos limitados.

1 Introdução

Modelos de Língua de Grande Porte (LLMs) desencadearam uma revolução em diversas áreas do conhecimento (Bubeck et al., 2023). No domínio da saúde, esses modelos têm demonstrado capacidade relevante para apoiar tarefas complexas, como a sumarização de prontuários médicos (Schneider et al., 2025), a aceleração da produção científica (Aygün et al., 2025) e o auxílio ao diagnóstico (Shan et al., 2025). Esse cenário também tem impulsionado o desenvolvimento de modelos especializados, como o MedGemma (Sellergren et al., 2025).

Apesar disso, a maioria desses modelos especializados é treinada predominantemente em inglês, o que impõe desafios para idiomas com menor representação, como o português (Bommasani et al., 2021; Chang et al., 2024). Embora existam modelos treinados para tarefas médicas em português, eles tendem a empregar arquiteturas mais antigas (Schneider et al., 2021) ou a focar em tarefas específicas (Schneider et al., 2025), não alcançando o caráter generalista e o escopo de conhecimento de modelos como o MedGemma, por exemplo. Assim, permanece em aberto se as capacidades multilíngues declaradas desses sistemas de maior porte são suficientes para garantir um desempenho confiável em contextos médicos, em que a precisão terminológica é crítica, fora do inglês.

Nesse contexto, a Geração Aumentada por Recuperação (RAG) (Lewis et al., 2020) surge como uma abordagem relevante para ampliar as capacidades dos LLMs, especialmente em domínios especializados e sensíveis como o médico. Ao permitir a consulta a bases externas de conhecimento durante a inferência, o RAG ancora as respostas em informações contextuais e, potencialmente, mais confiáveis, reduzindo a dependência exclusiva do conhecimento paramétrico do modelo (Izacard e Grave, 2021). Dessa forma, a técnica pode contribuir para mitigar limitações linguísticas e lacunas de conhecimento associadas a idiomas menos representados, oferecendo um mecanismo complementar para lidar com variações de idioma sem a necessidade de retreinamento extensivo do modelo.

Desse modo, este trabalho investiga o impacto

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 467–477

da variação linguística e do uso de recuperação de conhecimento externo no desempenho de LLMs em tarefas médicas. Com tal objetivo, pretende-se responder às seguintes questões de pesquisa (QPs). A **QP1.** *"Em que medida a variação do idioma de entrada impacta o desempenho dos LLMs em tarefas médicas?"* orienta a análise do efeito do idioma de entrada sobre o desempenho dos modelos, enquanto a **QP2.** *"A integração de uma base de conhecimento externa em português via RAG mitiga vieses linguísticos e melhora o desempenho dos modelos?"* investiga o papel da recuperação de conhecimento externo como mecanismo de mitigação dos efeitos da variação linguística. Adicionalmente, a **QP3.** *"Essa variação afeta o processo de raciocínio explícito dos LLMs?"* complementa a análise ao focar na influência da língua de entrada sobre o raciocínio.

Para responder a essas questões, realizamos experimentos em português e em inglês com duas variantes do MEDGEMMA (Schneider et al., 2025), de diferentes portes (4B e 27B), utilizando três conjuntos de dados — medical-o1-reasoning (Chen et al., 2025), MedQA (Jin et al., 2021) e Global-MMLU (Singh et al., 2025) — com conteúdo médico e avaliando também uma configuração com RAG. Os resultados indicam que a variabilidade linguística tende a diminuir à medida que a escala do modelo aumenta, enquanto a variante 4B apresenta desempenho consistentemente inferior em português. A incorporação de RAG, avaliada no modelo de menor porte, não foi suficiente para compensar essa diferença, sugerindo limitações do modelo para explorar de forma eficaz o volume adicional de informação fornecido. Em conjunto, os achados evidenciam a interação entre a escala do modelo, a variação linguística e o uso de RAG, contribuindo para uma compreensão mais precisa dos desafios na aplicação de LLMs médicos em português.

2 Trabalhos Relacionados

Esta seção revisa brevemente trabalhos relacionados à aplicação de RAG no domínio médico e ao desenvolvimento de sistemas generativos de apoio no domínio da saúde.

A técnica RAG tem se destacado em diversas aplicações por integrar modelos de língua a fontes externas de conhecimento, ampliando a precisão factual e a confiabilidade das respostas (Arslan et al., 2024; Abo El-Enen et al., 2025). Trabalhos recentes indicam seu potencial para reduzir alucinações e aumentar a explicabilidade de sistemas generativos, particularmente em domínios sensíveis como o médico (Yang et al., 2025).

Nesse mesmo domínio, trabalhos também investigam *chatbots* médicos baseados em LLMs como ferramentas de apoio. Essas abordagens demonstram potencial para automatizar tarefas, como a identificação de medicamentos, a verificação de interações e o fornecimento de orientações iniciais a pacientes (Kim et al., 2024). O *Robotic Medical Support Chatbot* (S et al., 2024), por exemplo, adota um modelo supervisionado com dois componentes, classificação da consulta em categorias de doenças e geração de instruções gerais de primeiros socorros. Embora não utilize LLMs, representa uma alternativa complementar. Ainda assim, tais sistemas enfrentam limitações de confiabilidade nas respostas e são frequentemente propostos para o inglês.

No contexto da língua portuguesa, iniciativas recentes buscam mitigar essas limitações. Destaca-se um sistema de perguntas e respostas em português brasileiro, baseado em RAG para consulta de bulas de medicamentos (Navarro et al., 2025). O estudo propõe um arcabouço de avaliação que separa estritamente a base de conhecimento da RAG do conjunto de perguntas de avaliação, evitando a sobreposição textual e garantindo uma avaliação centrada na recuperação e na síntese. Embora não replicada, essa lógica é semelhante à adotada neste trabalho, ao filtrar um dos conjuntos de avaliação para manter apenas entradas com baixa similaridade entre si na base de conhecimento do RAG (Seção 3.1.1).

Por fim, embora existam esforços para a construção de *benchmarks* multilíngues no domínio médico, como o MedExpQA (Alonso et al., 2024), o português permanece significativamente menos explorado nessas avaliações. Essa lacuna limita a compreensão do desempenho de modelos em cenários médicos mais gerais, como o clínico, fora do eixo anglófono. Este trabalho busca contribuir para mitigar essa ausência ao avaliar sistematicamente o impacto da língua de inferência no desempenho de sistemas RAG em tarefas médicas, analisando como escolhas linguísticas influenciam tanto a recuperação de informações quanto a qualidade das respostas geradas.

3 Metodologia

Nesta seção, apresentamos a metodologia experimental adotada, incluindo a preparação dos dados para a realização dos experimentos, em português e em inglês, a modelagem experimental, as métricas de avaliação e os detalhes de implementação.

3.1 Preparação dos Dados

Esta seção descreve os processos de preparação, filtragem e tradução dos conjuntos de dados utilizados.

3.1.1 Conjuntos de Dados e Particionamento

Três conjuntos de dados foram utilizados neste estudo: **(a) medical-o1-reasoning** (Chen et al., 2025), composto por 19,7 mil instâncias de raciocínio médico derivadas de problemas verificáveis e validadas por um verificador baseado em LLM. Cada entrada contém a descrição do problema, uma pergunta, majoritariamente baseada em casos médicos, o raciocínio e a resposta verificada do GPT-4o (OpenAI et al., 2024); **(b) MedQA** (Jin et al., 2021), um conjunto de 10,1 mil perguntas de múltipla escolha, predominantemente voltadas a diagnóstico, coletadas a partir de exames profissionais de conselhos médicos; e **(c) Global-MMLU** (Singh et al., 2025), um *benchmark* multilíngue que inclui inglês e português, contendo cerca de 1,5 mil questões de múltipla escolha da área médica.

Para os experimentos, foi necessário um pré-processamento dos conjuntos. Observou-se que muitas instâncias de **(a)** eram similares às de **(b)**; assim, removemos de **(b)** as entradas com similaridade de cosseno superior a 0,75, resultando em 5,3 mil instâncias. Além disso, como **(a)** e **(b)** estavam disponíveis apenas em inglês, realizamos tradução automática para o português para atender ao nosso objetivo de avaliar o impacto do idioma de entrada. Esse procedimento não foi necessário para **(c)**, que já contempla ambos os idiomas.

Os conjuntos **(a)** e **(b)** — com exceção de 500 instâncias de **(b)**, que serviram para avaliação — foram utilizados como base de conhecimento do sistema RAG, enquanto **(c)** foi utilizado exclusivamente para avaliação. Essa escolha deve-se ao fato de **(b)** e **(c)** conterem questões de múltipla escolha, o que facilita a avaliação em larga escala, e à disponibilidade, em **(c)**, de pares já nos idiomas-alvo, o que proporciona um conjunto de avaliação robusto.

3.1.2 Tradução dos Conjuntos de Dados

Para a tradução dos conjuntos de dados, considerando os recursos disponíveis, foram avaliados o modelo geral Llama-3.1 8B (AI@Meta, 2024) e o modelo especializado em tradução GemmaX2-28-9B (Cui et al., 2025), que apresenta resultados no estado da arte. Ambos foram investigados devido à natureza altamente especializada do domínio, uma vez que não era evidente se um modelo focado em tradução lidaria adequadamente com terminologia médica.

As traduções foram avaliadas por meio de um conjunto robusto de métricas, incluindo as métricas léxicas BLEU (Papineni et al., 2002) e ROUGE-L (Lin, 2004) (F1), a métrica semântica BERTScore (Zhang et al., 2020) (F1), e as métricas específicas de tradução Comet (Rei et al., 2020), COMETKiwi (Rei et al., 2023) e xCOMET (Guerreiro et al., 2024). Com exceção do COMETKiwi, todas as métricas requerem referência. Assim, adotou-se um processo de retrotradução (Sennrich et al., 2016), no qual os textos em inglês foram traduzidos para o português e, em seguida, traduzidos novamente para o inglês. A comparação é então realizada entre os textos em inglês, com a intuição de que, se a tradução for acurada, o conteúdo será preservado. A Tabela 1 apresenta os resultados, indicando um desempenho consistentemente superior do GemmaX2, cujas traduções foram adotadas nos experimentos subsequentes.

Métrica	Llama3.1 8B	GemmaX2-28-9B
BLEU	42,699	**50,397**
ROUGE-L	0,620	**0,748**
BERTScore	0,910	**0,956**
COMET	0,782	**0,827**
COMETKiwi	0,717	**0,801**
xCOMET	0,498	**0,772**

Tabela 1: Avaliação da Tradução. Melhores resultados em **negrito**.

3.2 Cenários de Avaliação

Para responder à QP1, consideramos diferentes cenários de avaliação. Utilizamos dois LLMs médicos, especificamente as variantes 4B e 27B do modelo MEDGEMMA (Schneider et al., 2025), citadas como referências em tarefas generativas no domínio médico. Esses modelos apresentam capacidades multilíngues, aspecto que viabiliza a análise do impacto da variação inglês–português.

Para cada conjunto de dados avaliado, aplicamos

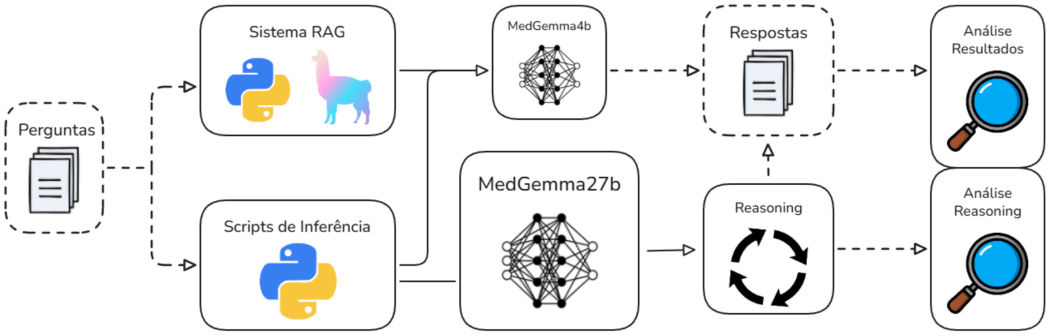


Figura 1: Visão geral das configurações experimentais avaliadas.

os procedimentos ilustrados na Figura 1. Na variante 4B, os experimentos foram conduzidos por meio de chamadas via *script* em Python, considerando entradas em português e em inglês, bem como uma configuração com RAG[1] em português, permitindo analisar o efeito do idioma e do uso de RAG no desempenho. Para a variante 27B, realizamos chamadas diretas, também via script em Python, em ambos os idiomas, avaliando se o modelo de maior porte é mais robusto à mudança de língua. Além disso, essa variante explicita sua cadeia de raciocínio antes da resposta final, permitindo investigar como o idioma de entrada influencia o processo de raciocínio.

Nesse contexto, a QP2 é avaliada na configuração com a variante 4B, por representar um cenário mais sensível — dado o tamanho do modelo — à variação linguística, no qual o uso de RAG em português pode desempenhar um papel mais relevante. Reconhecemos que uma avaliação comparativa do RAG na variante de maior porte também seria pertinente, contudo, limitações de recursos computacionais impediram sua realização (detalhes na Seção 3.4). Por outro lado, a QP3 é analisada com base na variante MedGemma 27B, que explicita sua cadeia de raciocínio e permite examinar a influência do idioma sobre o processo inferencial.

3.3 Métricas de Avaliação

A avaliação dos resultados foi estruturada em duas frentes distintas, alinhadas aos objetivos deste trabalho. A primeira frente concentra-se na quantificação do impacto do uso de RAG no desempenho geral dos modelos, enquanto a segunda explora, de forma qualitativa e quantitativa, os efeitos da variação linguística tanto no processo de raciocínio quanto nos resultados obtidos em diferentes cenários de execução.

[1]Instanciado usando o *framework* LlamaIndex (Liu, 2022).

A **acurácia** foi adotada como a métrica principal para a avaliação do desempenho global em todos os cenários experimentais, correspondendo à proporção de respostas corretas produzidas pelos modelos nos conjuntos de dados de avaliação. Complementarmente, conduzimos uma análise quantitativa dos passos lógicos explícitos gerados pelo modelo MedGemma 27B, com foco no seu processo de raciocínio. Essa análise considerou o **número médio de passos de raciocínio** obtido a partir de entradas de perguntas em inglês e em português. A contagem foi automatizada por meio da expressão regular `(?m)^\s*\d+\.`, que identifica itens numerados em listas estruturadas conforme o padrão de formatação das respostas do modelo. A métrica foi calculada como a diferença entre o número de passos gerados a partir da pergunta em inglês e o gerado a partir da pergunta em português.

Também analisamos o **tamanho médio dos textos de raciocínio**, medido em número de caracteres, produzidos pelo MedGemma 27B para ambos os idiomas de entrada. Consideramos também a **preservação semântica do raciocínio** entre inglês e português, avaliada por meio do BERTScore, aplicado à comparação de textos gerados a partir da mesma pergunta, em idiomas distintos.

Além disso, investigamos a sobreposição de erros entre idiomas em um mesmo modelo, visando identificar se os erros tendem a ocorrer nas mesmas questões, independentemente da língua de entrada. Para isso, calculamos a proporção de erros em comum entre as respostas em português e em inglês. Formalmente, considerando L_{pt} e L_{en} como os conjuntos de respostas em português e inglês, respectivamente, L_{ref} como a língua de referência do cálculo e $E(L)$ como o subconjunto de respostas incorretas em um conjunto L, a métrica é

$$\text{Valor} = \frac{|E(L_{pt}) \cap E(L_{en})|}{|E(L_{ref})|}.$$

Por fim, avaliamos o **impacto da qualidade da tradução** no desempenho dos modelos por meio da análise da acurácia em subconjuntos de perguntas do MedQA. As perguntas foram estratificadas nas cem melhores e nas cem piores traduções, segundo as métricas COMETKiwi e xCOMET, o que permitiu examinar a influência direta da qualidade da tradução da pergunta sobre o resultado. Essas métricas foram escolhidas por apresentarem características complementares, uma vez que o COMETKiwi é livre de referência, enquanto o xCOMET é reconhecido como estado da arte acadêmico na avaliação de traduções (Guerreiro et al., 2024).

3.4 Configurações Experimentais

Os experimentos foram conduzidos com duas GPUs NVIDIA RTX 4090. Para viabilizar a execução da variante MedGemma 27B em um ambiente com limitações de memória, o modelo foi carregado com quantização de 8 bits (Lang et al., 2024), uma abordagem que permite reduzir significativamente o uso de VRAM, com pouco impacto na qualidade da inferência (Huang et al., 2024). O parâmetro de geração `max_new_tokens` foi ajustado conforme a variante do modelo, definido como 500 para a variante 4B e 1500 para a variante 27B, de modo a acomodar a geração completa do texto de raciocínio no último caso. Em todos os cenários, o parâmetro `do_sample` foi fixado em `False`, o que proporcionou maior reprodutibilidade experimental.

O agente RAG foi instanciado por meio do *framework* LlamaIndex (Liu, 2022). A construção do banco de dados vetorial explorou a estrutura inerente dos dados, dessa forma cada bloco de informação (*chunk*) inserido no índice vetorial foi formatado a partir da concatenação de um par de pergunta e resposta dos conjuntos de dados. Os *embeddings* foram gerados pelo modelo MULTILINGUAL-E5-LARGE (Wang et al., 2024), escolhido não apenas por sua capacidade multilíngue, mas também por apresentar um dos melhores desempenhos no *benchmark* de referência MTEB (Muennighoff et al., 2023), mantendo exigências computacionais compatíveis com os nossos recursos.

Para a etapa de recuperação, adotou-se `top_k = 2`. A escolha restrita desse valor foi motivada pela utilização exclusiva da variante menor do MedGemma, com "apenas" 4 bilhões de parâmetros, no pipeline, equilibrando a necessidade de acesso à informação relevante e a mitigação do efeito de *lost in the middle* (Liu et al., 2024). Essa configuração favorece a precisão do contexto fornecido ao modelo gerador, especialmente em cenários com modelos de menor porte. Por fim, os documentos recuperados são incorporados ao contexto do *prompt*, ilustrado na Figura 2.

Prompt

«CONTEXT»{context_text}«END_CONTEXT»

Responda com uma das opções entre colchetes (`[A]`, `[B]`, `[C]` ou `[D]`).

Escreva **APENAS** a **LETRA** da resposta (`[A]`, `[B]`, `[C]` ou `[D]`) entre `<answer></answer>`.

{question_text}

Figura 2: *Prompt* usado para responder às perguntas.

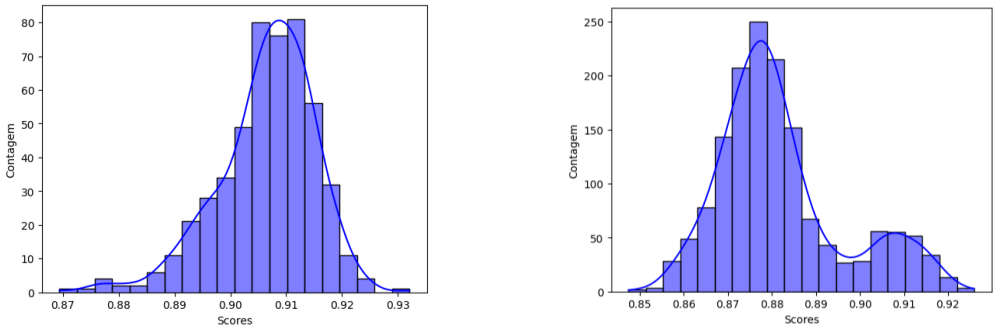
4 Resultados

Esta seção apresenta a análise dos resultados. Primeiro, a Tabela 2 fornece evidências diretamente relacionadas à QP1, ao permitir analisar como a variação do idioma de entrada afeta o desempenho dos modelos. De forma geral, observa-se que a variante MedGemma 27B apresenta acurácia consistentemente superior à da variante 4B em todos os cenários avaliados. Além disso, o impacto do idioma não é uniforme, variando conforme o modelo e o conjunto de dados considerados.

Conjunto	Modelo	EN (s/RAG)	PT (s/RAG)
Global-MMLU	MedGemma-4B	**64,65**	57,48
Global-MMLU	MedGemma-27B	81,66	**81,79**
MedQA	MedGemma-4B	**59,00**	42,80
MedQA	MedGemma-27B	72,80	**77,20**
Ablação do RAG (MedGemma-4B)			
Global-MMLU	MedGemma-4B	**PT c/RAG**	55,55
MedQA	MedGemma-4B	**PT c/RAG**	43,80

Tabela 2: Resultados de acurácia (%). Melhores resultados *por modelo* em **negrito**.

No caso do MEDGEMMA 27B, o desempenho permanece elevado e amplamente estável entre inglês e português no Global-MMLU, com, inclusive, um pequeno ganho de acurácia no português no conjunto MedQA. Em contraste, o MEDGEMMA 4B sem RAG apresenta uma redução sistemática de acurácia ao operar em português nos dois



(a) MEDQA (b) GLOBALMMLU

Figura 3: Distribuição da média dos *scores* de recuperação.

	MedQA				**GlobalMMLU**			
	MedGemma-4b		**MedGemma-27b**		**MedGemma-4b**		**MedGemma-27b**	
	%	**Sup.**	**%**	**Sup.**	**%**	**Sup.**	**%**	**Sup.**
Português	61,19	286	62,28	114	68,91	640	76,64	274
Inglês	85,37	205	52,21	136	82,89	532	76,09	276

Tabela 3: Proporção e suporte da sobreposição de erros entre os idiomas.

conjuntos de dados, indicando maior sensibilidade linguística em modelos de menor porte.

No geral, os resultados indicam que a robustez multilíngue tende a aumentar à medida que a escala do modelo aumenta. Na variante 27B, a proximidade de desempenho entre idiomas sugere representações internas e capacidades de raciocínio mais transferíveis entre inglês e português, possivelmente decorrentes de melhores representações multilíngues e de maior capacidade paramétrica, que permite maior espaço para codificação de informações. Por outro lado, a queda consistente observada no modelo de menor porte aponta para limitações na representação do português, refletindo-se em acurácia significativamente inferior nesse idioma.

A Tabela 2 também permite analisar a QP2, avaliando se a incorporação de RAG melhora o desempenho em português. A adição de contexto recuperado ao MEDGEMMA 4B não foi suficiente para aproximar o desempenho em português do observado em inglês. De modo geral, o uso de RAG não mitiga a perda de acurácia associada à variação linguística. Inclusive, no Global-MMLU, o desempenho com RAG é ainda inferior, enquanto, no MedQA, observa-se apenas uma melhora muito modesta.

Apesar disso, observa-se que a qualidade da recuperação dos documentos foi elevada na maioria das perguntas. Para investigar se o baixo desempenho estava associado à etapa de recuperação, a Figura 3 apresenta as distribuições dos *scores* médios de similaridade entre cada pergunta e o contexto recuperado. Tanto no MedQA quanto no Global-MMLU, a maioria das consultas atingiu valores superiores a 0,85, com picos próximos de 0,88 e 0,91, respectivamente. Ainda assim, essa elevada similaridade não se traduziu em ganhos consistentes de desempenho. Esses resultados indicam que o principal ponto crítico do processo não reside na etapa de busca, mas sim na forma como o modelo explora o contexto recuperado. Em particular, o MEDGEMMA 4B, devido à sua capacidade paramétrica mais limitada, pode enfrentar dificuldades para filtrar informações redundantes ou lidar com *prompts* mais extensos, que demandam maior capacidade de memória contextual. Dessa forma, mesmo quando o RAG fornece passagens altamente relevantes, o modelo nem sempre consegue transformá-las em melhorias na geração final.

A Tabela 3 aprofunda a análise do impacto do idioma sobre o desempenho dos modelos ao apresentar a proporção de sobreposição de erros entre inglês e português para cada combinação de modelo e conjunto de dados. No MedQA, observa-se uma assimetria pronunciada em favor do MEDGEMMA 4B, com alta sobreposição quando o inglês é tomado como referência (85,37%) e um valor substancialmente menor no português (61,19%). Esse padrão não se repete na variante MEDGEMMA

27B, na qual os valores entre os idiomas se aproximam e a sobreposição no português aumenta. No Global-MMLU, o comportamento da variante 4B permanece semelhante ao observado no MedQA, enquanto os resultados da variante 27B apresentam maior convergência.

Uma alta sobreposição indica que erros cometidos em uma língua tendem a se repetir na outra, sugerindo limitações intrínsecas do modelo em relação às questões em si. Nesse sentido, os valores significativamente menores de sobreposição para o português no MEDGEMMA 4B indicam que a variação linguística afeta negativamente sua capacidade de resposta. Em contraste, a maior proximidade dos valores observada no MEDGEMMA 27B reforça a evidência de que modelos de maior escala são menos sensíveis à variação linguística.

Em relação à QP3, que investiga como a variação português–inglês na entrada do modelo afeta o processo de raciocínio, a Tabela 4 apresenta o número médio de passos lógicos gerados pelo MEDGEMMA 27B em ambos os idiomas e os conjuntos de dados. Ressalta-se, novamente, que essa é a única variante avaliada que oferece suporte nativo à explicitação do raciocínio. Observa-se que as respostas em português apresentam, em média, mais passos lógicos do que as geradas a partir das mesmas perguntas em inglês.

MedQA		**GlobalMMLU**	
Português	**Inglês**	**Português**	**Inglês**
6,1	5,5	7,1	6,7

Tabela 4: Número médio de passos lógicos para o modelo MedGemma 27B.

Para aprofundar essa análise, calculamos a diferença média no número de passos entre os raciocínios produzidos a partir das versões em inglês e em português de cada pergunta. As diferenças médias obtidas foram de $-0{,}51$ no MedQA e de $-0{,}36$ no Global-MMLU, confirmando uma tendência consistente de maior detalhamento do raciocínio em português. No entanto, a baixa magnitude dessas diferenças indica que, apesar do aumento no número de passos, a estrutura global do raciocínio permanece estável entre os idiomas.

Durante a inspeção qualitativa das respostas, observamos que, quando a entrada é em português, o modelo tende a traduzir explicitamente os conceitos apresentados nas alternativas como parte do processo de avaliação, como no exemplo:

> "*Esferas cheias de endósporos (Spherules filled with endospores)* [...]"

Esses trechos correspondem, em geral, aos únicos momentos em que o português aparece no raciocínio explícito do modelo. Tal comportamento é esperado, dado que o treinamento dessa habilidade nos modelos ocorre predominantemente em inglês. Assim, mesmo com perguntas e respostas finais em português, o processo de raciocínio permanece majoritariamente no outro idioma.

MedQA		**GlobalMMLU**	
Português	**Inglês**	**Português**	**Inglês**
4249,3	3994,4	3085,1	2921,0

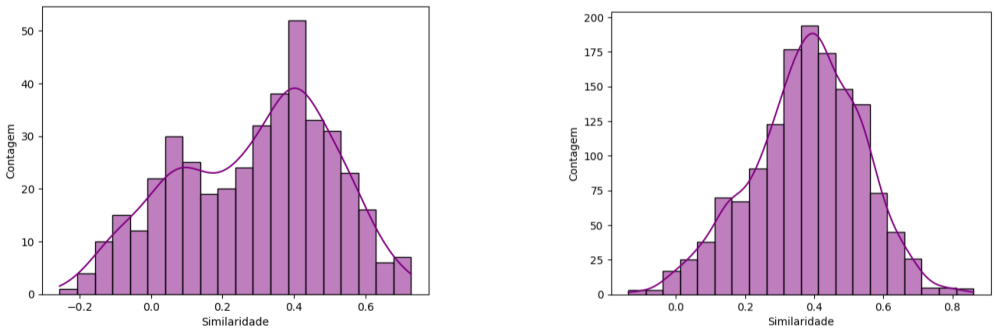
Tabela 5: Média de caracteres do texto de raciocínio.

Outra característica recorrente quando as entradas são em português é a presença de um passo adicional de verificação ao final do raciocínio. Esse passo ocorre após a etapa de conclusão e consiste em uma checagem explícita, como no exemplo:

> "*6. **Conclusion:** The clinical presentation is highly suggestive of coccidioidomycosis (Valley Fever). The characteristic finding of *Coccidioides* in tissue biopsy is spherules containing endospores. Therefore, option [A] is the answer.*
>
> *7. **Final check:** The patient's symptoms [...] and history [...] strongly point to coccidioidomycosis. The microscopic description in [A] perfectly matches the tissue form of *Coccidioides*.*"

Esse padrão não é observado com a mesma frequência nas entradas em inglês, nas quais o raciocínio tende a se encerrar diretamente na etapa de conclusão. Esse comportamento sugere que o modelo apresenta maior incerteza quando a inferência é realizada em português, recorrendo a etapas adicionais de verificação para compensar a incerteza linguística. Como consequência, os raciocínios produzidos em português tendem a ser mais extensos do que os gerados em inglês, conforme os resultados da Tabela 5.

A Figura 4 complementa essa análise ao considerar a dimensão semântica, por meio do cálculo do BERTScore entre os textos de raciocínio gerados para a mesma pergunta em inglês e em português, avaliando em que medida o conteúdo semântico é preservado. As distribuições revelam variabilidade



(a) MEDQA (b) GLOBALMMLU

Figura 4: Distribuição de similaridades entre os raciocínios português–inglês.

	COMETKiwi		xCOMET	
Modelo	**100 Piores**	**100 Melhores**	**100 Piores**	**100 Melhores**
MedGemma-4b	39,00%	**47,00%**	35,00%	**53,00%**
MedGemma-27b	**75,00%**	63,00%	56,00%	**86,00%**
RAG (4b)	**44,00%**	39,00%	31,00%	**50,00%**

Tabela 6: Acurácia por Faixa de Qualidade da Tradução. Melhores resultados por modelo em **negrito**.

entre os conjuntos de dados, com uma concentração de casos de alta similaridade e uma cauda relevante de baixa similaridade, mais acentuada no MedQA (Figura 4a). Esses resultados indicam que a preservação semântica do raciocínio nem sempre é mantida ao traduzir a pergunta, especialmente no MedQA, conjunto no qual o modelo apresentou desempenho inferior em termos de acurácia, sugerindo que dificuldades na resolução da tarefa estão associadas a maiores divergências semânticas no raciocínio entre idiomas.

Por fim, investigamos se efeitos introduzidos pela tradução podem atuar como fonte de ruído e afetar o desempenho dos modelos. Para isso, a Tabela 6 apresenta as acurácias obtidas em subconjuntos compostos pelas 100 melhores e pelas 100 piores traduções, segundo as métricas COMETKiwi e xCOMET. Considerando o xCOMET, observa-se que as acurácias são consistentemente mais altas no subconjunto das melhores traduções, reforçando a interpretação de que a tradução para o português pode introduzir fatores que afetam negativamente o desempenho final do sistema, especialmente em um domínio altamente especializado, no qual a tradução permanece um desafio em aberto.

Em contraste, quando avaliada pelo COMETKiwi, observa-se uma redução inesperada de acurácia no subconjunto das melhores traduções para o MedGemma 27B e para o sistema com RAG. Essa discrepância entre as métricas sugere que determinadas traduções podem ter perdido parte do conteúdo durante o processo de tradução, o que afeta os modelos. Embora esses resultados indiquem que a tradução não é isenta de impacto no desempenho dos modelos e represente uma direção relevante para investigações futuras, os resultados apresentados anteriormente indicam que a parte crítica das diferenças observadas decorre do idioma de entrada, o que se reflete em comportamentos distintos entre os modelos.

5 Conclusões

Este trabalho avaliou os efeitos da variação do idioma de entrada português–inglês em LLMs especializados no domínio médico. Para isso, foram conduzidos experimentos com duas variantes do modelo MedGemma (4B e 27B parâmetros), utilizando três conjuntos de dados médicos e combinando avaliações quantitativas de desempenho com análises estruturais das gerações. Além disso, investigaram-se o impacto da técnica de RAG no modelo de menor porte e os efeitos da variação linguística sobre o raciocínio explícito do modelo de maior escala.

Os resultados indicam que a robustez ao idioma é significativamente maior na variante com 27B, enquanto o modelo menor apresenta desempenho inferior em português. Embora o sistema de RAG tenha demonstrado alta qualidade na recuperação

de contexto relevante, sua incorporação não foi suficiente para compensar a perda de desempenho do modelo menor ao operar em português, sugerindo limitações do modelo menor na exploração eficaz de contextos extensos e fora do inglês.

Entre os trabalhos futuros, destacam-se a construção de recursos médicos nativos em português, a análise sistemática das gerações por especialistas e a ampliação do escopo experimental, incluindo outras configurações de RAG, hiperparâmetros e LLMs. Ainda assim, este trabalho contribui para o entendimento dos limites e desafios da aplicação de LLMs médicos em idiomas além do inglês, com foco especial no português.

Limitações

A análise baseou-se em recursos médicos traduzidos automaticamente para viabilizar a comparação entre idiomas, dada a escassez de recursos estruturados em português. Esse processo pode introduzir ruído e constituir uma limitação potencial. Assim, embora as métricas automáticas selecionadas sejam adequadas, podem não captar plenamente as nuances semânticas e as especificidades do domínio. Uma validação em pequena escala, conduzida por um especialista em saúde, poderia fornecer evidências mais robustas sobre a preservação do significado original. Da mesma forma, o desempenho em tarefas de recuperação semântica deveria ser complementado por avaliação especializada, em vez de depender exclusivamente de pontuações médias de similaridade, que podem superestimar a adequação. A validação por especialistas aumentaria, portanto, a confiabilidade dos resultados obtidos tanto na tradução quanto na recuperação.

Declaração Ética

Este trabalho reconhece as implicações éticas do uso de modelos generativos no contexto médico. As abordagens investigadas têm caráter estritamente acadêmico e experimental e não devem, em nenhuma circunstância, ser interpretadas como substitutas do julgamento, da responsabilidade ou da atuação de profissionais de saúde qualificados. Sistemas baseados em LLMs podem atuar como ferramentas de apoio informacional ou assistencial, contribuindo para a organização, análise ou contextualização de informações médicas, mas não devem ser utilizados de forma *totalmente* autônoma para diagnóstico, prescrição ou tomada de decisões clínicas. Ressaltamos ainda a importância da validação rigorosa, da supervisão humana e do uso responsável dessas tecnologias.

Agradecimentos

Esta pesquisa foi financiada pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), código de financiamento 307088/2023-5, pela FAPERJ – Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, processos SEI-260003/002930/2024 e SEI-260003/000614/2023, e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. O trabalho também contou com o apoio dos Institutos Nacionais de Ciência e Tecnologia (INCTs) do CNPq, IAIA (grant 406417/2022-9), TILD-IAR (grant 408490/2024-1) e IAPROBEM (grant 408589/2024-8).

Referências

Mohamed Abo El-Enen, Sally Saad, e Taymoor Nazmy. 2025. A survey on retrieval-augmentation generation (rag) models for healthcare applications. *Neural Computing and Applications*, 37(33):28191–28267.

AI@Meta. 2024. Llama 3 model card.

Iñigo Alonso, Maite Oronoz, e Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, e Christophe Cruz. 2024. A survey on rag with llms. *Procedia computer science*, 246:3781–3790.

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, Marc Coram, Hao Cui, Jake Garrison, Renee Johnston Anton Kast, Cory Y. McLean, Peter Norgaard, Zahra Shamsi, David Smalling, James Thompson, Subhashini Venugopalan, Brian P. Williams, Chujun He, Sarah Martinson, Martyna Plomecka, Lai Wei, Yuchen Zhou, and 23 others. 2025. An ai system to help scientists write expert-level empirical software. *Preprint*, arXiv:2509.06503.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, and 95 others. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro,

e Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, e Ben Bergen. 2024. When is multilinguality a curse? language modeling for 250 high-and low-resource languages. Em *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, páginas 4074–4096.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, e Benyou Wang. 2025. Towards medical complex reasoning with llms through medical verifiable problems. Em *Findings of the Association for Computational Linguistics: ACL 2025*, páginas 14552–14573, Vienna, Austria. Association for Computational Linguistics.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, e Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. Em *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, páginas 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, e André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, e Xiaojuan Qi. 2024. Billm: pushing the limit of post-training quantization for llms. Em *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Gautier Izacard e Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. Em *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, páginas 874–880.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, e Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Won Tae Kim, Jaegwang Shin, In-Sang Yoo, Jae-Woo Lee, Hyun Jeong Jeon, Hyo-Sun Yoo, Yongwhan Kim, Jeong-Min Jo, ShinJi Hwang, Woo-Jeong Lee, Seung Park, e Yong-June Kim. 2024. Medication extraction and drug interaction chatbot: Generative pre-trained transformer-powered chatbot for drug-drug interaction. *Mayo Clinic Proceedings: Digital Health*, 2(4):611–619.

Jiedong Lang, Zhehao Guo, e Shuyu Huang. 2024. A comprehensive study on quantization techniques for large language models. Em *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, páginas 224–231.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, e Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Em *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. Em *Text Summarization Branches Out*, páginas 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jerry Liu. 2022. LlamaIndex.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, e Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, e Nils Reimers. 2023. MTEB: Massive text embedding benchmark. Em *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Letícia C. Navarro, Filipe Mutz, Thiago M. Paixão, Guilherme G. Zanetti, Claudine Badue, Alberto F. De Souza, e Thiago Oliveira-Santos. 2025. Ragpharma: A rag-based chatbot for medicine leaflets with a dual-dataset evaluation framework. *Journal of the Brazilian Computer Society*, 31(1):1137–1149.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, e Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. Em *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, página 311–318, USA. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, e André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. Em *Proceedings of the Eighth Conference on Machine Translation*,

páginas 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, e Alon Lavie. 2020. COMET: A neural framework for MT evaluation. Em *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 2685–2702, Online. Association for Computational Linguistics.

Sreedhar Kumar S, Syed Thouheed Ahmed, Afifa Salsabil Fathima, Nishabai M, e Sophia S. 2024. Medical chatbot assistance for primary clinical guidance using machine learning techniques. *Procedia Computer Science*, 233:279–287. 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024).

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Yohan Bonescki Gumiel, Claudia Moro, e Emerson Cabrera Paraiso. 2021. A gpt-2 language model for biomedical texts in portuguese. Em *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, páginas 474–479.

Elisa Terumi Rubel Schneider, Fernando Henrique Schneider, Emerson Cabrera Paraiso, Alceu Souza Britto Jr, e Rafael Menelau Oliveira Cruz. 2025. Medgemma-sum-pt: A lightweight model for portuguese clinical summarization. Em *CLEF 2025 Working Notes: Notebook for the BioASQ Lab at CLEF 2025*, Madrid, Spain. CEUR Workshop Proceedings. Available under Creative Commons Attribution 4.0 International (CC BY 4.0).

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Rico Sennrich, Barry Haddow, e Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. Em *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 86–96, Berlin, Germany. Association for Computational Linguistics.

Guxue Shan, Xiaonan Chen, Chen Wang, Li Liu, Yuanjing Gu, Huiping Jiang, e Tingqi Shi. 2025. Comparing diagnostic accuracy of clinical professionals and large language models: Systematic review and meta-analysis. *JMIR medical informatics*, 13:e64963.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. Em *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, e Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle Bitterman, Jasmine Ong, Daniel Ting, e Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, e Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. Em *International Conference on Learning Representations*.