Avaliação *End-to-End* de um Sistema RAG para Documentos Hospitalares em Português

Murilo Vargas da Cunha[1,2] **Marília Rosa Silveira**[1] **César Brasil Sperb**[1]
Brenda Salenave Santana[1] **Larissa Astrogildo Freitas**[1] **Ulisses Brisolara Corrêa**[1]
[1]Universidade Federal de Pelotas (UFPEL), Pelotas, RS, Brasil
[2]Instituto Federal do Rio Grande do Sul (IFRS), Rio Grande, RS, Brasil
{mvcunha, mrsilveira, cbsperb, bssalenave, larissa, ulisses}@inf.ufpel.edu.br

Resumo

Este artigo avalia um sistema *end-to-end* de Geração Aumentada por Recuperação (RAG) para consulta a documentos hospitalares regulatórios em português. O estudo analisa o impacto da otimização de cada componente (recuperação, reordenação e geração) em um cenário de recursos limitados. A metodologia combinou a criação de um *dataset* híbrido (sintético e validado por especialistas) com avaliações quantitativas utilizando métricas como MRR, NDCG@10 e BERTScore. Os resultados demonstram que o modelo de *embedding* `intfloat/multilingual-e5-small` apresentou a maior robustez, com taxa de falha de apenas 1,4% na recuperação. Na etapa de reclassificação, o método RRF destacou-se pelo equilíbrio entre custo computacional e desempenho. Conclui-se que a arquitetura otimizada, integrando esses componentes ao gerador Gemini 2.5 Flash, oferece uma solução eficiente e precisa para suporte à decisão em ambientes hospitalares.

1 Introdução

Interagir com sistemas conversacionais baseados em inteligência artificial se tornou comum em diversos serviços (Cederlund et al., 2024). Com o avanço dos Modelos de Linguagem Grandes (do inglês, *Large Language Models* – LLMs), a qualidade das respostas oferecidas por esses sistemas evoluiu de forma significativa (Raschka, 2024; Naveed et al., 2024). Na área da saúde, por exemplo, além de assistentes destinados a pacientes, surgem ferramentas voltadas a profissionais, pesquisadores e gestores, capazes de apoiar consultas rápidas a normas, protocolos e documentos institucionais (da Cunha et al., 2025). Por isso, abre-se a possibilidade de utilizar modelos de linguagem como instrumentos de apoio à tomada de decisão baseada em documentos para ambientes clínicos e administrativos.

Apesar desses avanços, a confiabilidade de LLMs em responder com base fiel em normas e procedimentos ainda é limitada: respostas podem omitir exceções, generalizar regras locais ou introduzir detalhes não presentes na documentação, o que é particularmente crítico em ambientes hospitalares e administrativos. Informações desatualizadas, interpretações equivocadas ou lacunas em áreas altamente especializadas afetam diretamente a segurança de uso desses sistemas em contextos sensíveis, como o setor hospitalar (Zhu et al., 2024; Abo El-Enen et al., 2025). Além disso, a maior parte das soluções disponíveis é voltada ao idioma inglês, criando barreiras para usuários e instituições baseadas em linguagens de poucos recursos, como o português (Alshammary et al., 2024; Chirkova et al., 2024).

Buscando superar essas limitações de cobertura e precisão dos LLMs, técnicas tradicionais de recuperação de informação, anteriormente empregadas principalmente em mecanismos de busca, foram integradas a modelos generativos, resultando em arquiteturas mais robustas e eficazes para leitura e compreensão de documentos (Raza et al., 2025). Essa abordagem é denominada Geração Aumentada por Recuperação (do inglês, *Retrieval-Augmented Generation* – RAG), que combina as capacidades generativas dos LLMs com mecanismos de recuperação de informação de bases de conhecimento externas (Ayala and Bechard, 2024; Oliveira et al., 2025).

Nesse cenário, a técnica de RAG se tornou uma estratégia para aumentar a precisão e a pertinência das respostas sem necessidade de retreinamento (Ayala and Bechard, 2024; Yu et al., 2025). Esta arquitetura permite que o modelo consulte documentos previamente determinados antes de formular a resposta, mitigando riscos de alucinação (Ji et al., 2023).

Estudos anteriores, realizaram avaliações quantitativas do componente de recuperação de um sis-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 749–758

tema RAG aplicado a documentos institucionais em português, analisando métricas como Precisão, Classificação Recíproca Média (do inglês, *Mean Reciprocal Rank* – MRR) e Ganho Cumulativo Descontado Normalizado (do inglês, *Normalized Discounted Cumulative Gain* – NDCG). Por outro lado, resta avaliar o desempenho final do *chatbot* sobre a união do componente de recuperação com a geração.

Neste trabalho, avançamos da avaliação de recuperação para uma avaliação *end-to-end*, o que inclui a etapa de geração de texto. As contribuições são: (i) ampliação do *benchmark*, incorporando 118 questões de uso real validadas por especialistas e totalizando 192 pares pergunta–resposta; (ii) comparação controlada de LLMs (via API e locais) mantendo o mesmo contexto recuperado; (iii) análise do *trade-off* entre custo e desempenho de reordenação, destacando quando métodos leves (tal como o RRF) se aproximam de alternativas baseadas em LLM.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia proposta, dividida em quatro etapas principais. A Seção 4 apresenta e discute os resultados obtidos. Finalmente, a Seção 5 apresenta as conclusões e delineia as direções para trabalhos futuros.

2 Trabalhos Relacionados

Embora a literatura sobre RAG esteja em expansão, a maioria das pesquisas concentra-se na língua inglesa (Ouyang, 2025). Dessa forma, nesta seção, revisamos estudos que exploram a aplicação de RAG em sistemas conversacionais voltados a línguas com poucos recursos.

O estudo de Obaid and Bawany (2024) apresenta o SeerahGPT, um sistema construído sobre a arquitetura Llama-2-7b que emprega RAG para recuperar dados de um corpus de textos islâmicos (Alcorão e Hadith). O trabalho aplica as métricas de Precisão Média (do inglês, *Mean Average Precision* – MAP) e Classificação Recíproca Média (do inglês, *Mean Reciprocal Rank* – MRR) para uma avaliação do sistema de recuperação e utiliza outras como Assistente de Avaliação Bilíngue (do inglês, *Bilingual Evaluation Understudy* – BLEU), ROUGE (do inglês, *Recall-Oriented Understudy for Gisting Evaluation*) e Métrica para Avaliação de Tradução com Ordenação Explícita (do inglês, *Metric for Evaluation of Translation with Explicit ORdering* – METEOR) para avaliar a geração de respostas. Wijaya and Purwarianti (2024) propõem um sistema de tutoria para programação, com base em documentos originalmente escritos no idioma indonésio, comparou diversos LLMs de código aberto, incluindo modelos das famílias Llama, Gemma, DeepSeek e Mistral, combinados com modelos de *embedding*, como e5-large-v2, jina-embeddings-v2 e gte-en-v1.5. A avaliação foi conduzida de forma qualitativa, além do uso da métrica MAP para avaliar a recuperação de documentos em diferentes tamanhos de *chunking*.

Seguindo essa mesma linha de avaliação da etapa de recuperação, o estudo de Nai et al. (2024) emprega o NDCG@10 para realizar uma extensa comparação entre múltiplos modelos de *embedding*, incluindo BM25, Cohere e vários modelos da OpenAI. Já o trabalho de Maryamah et al. (2024) usa outras métricas de ranqueamento e opta por uma avaliação baseada nas métricas clássicas de revocação e precisão para comparar seus modelos de *embedding*. Apesar da diversidade de métricas e de focos de análise, um ponto em comum entre estes trabalhos é a ausência de uma camada de reordenamento (*re-ranker*).

O reordenamento serve como refinamento dos resultados, de forma a otimizar o espaço de contexto de modelos, mantendo sempre os documentos que ajudam na resposta nas primeiras posições. O estudo proposto por Alonso et al. (2024), por exemplo, utiliza a Fusão de Classificação Recíproca (do inglês, *Reciprocal Rank Fusion* – RRF) para fundir e reordenar as listas de candidatos recuperados por múltiplas técnicas de recuperação da informação (BM25 e MedCPT), mas não avalia o impacto desse método na qualidade da recuperação. De forma análoga, no trabalho de Alshammary et al. (2024), os autores empregam um processo de filtragem e verificação de fatos (*"Fact Checking"*) que atua como um reordenador implícito, selecionando os documentos mais relevantes (*"Gold Hadiths"*) para a geração da resposta. Contudo, o estudo também não fornece uma análise quantitativa da eficácia dessa etapa de refino.

Em uma abordagem mais voltada para um contexto de estudos que exploraram documentos escritos em português, tem-se o trabalho de Seabra et al. (2024), em que os autores propuseram um sistema de *Question-Answering (Q&A)* voltado à gestão contratual, desenhado especificamente para processar o vernáculo jurídico e as obrigações contratuais brasileiras, integrando dados não estruturados (con-

tidos em arquivos do tipo PDF) e estruturados de sistemas de gestão. A metodologia inova ao empregar uma abordagem multiagente para combinar técnicas de RAG, utilizando modelos da série GPT, como o text-davinci-002 para *embeddings* e GPT-4-turbo para inferência. De forma análoga, Medeiros and de Oliveira (2025) realizaram uma análise comparativa de modelos de *embeddings* e LLMs no contexto RAG, com foco no desempenho dos módulos de recuperação e geração, em documentos escritos em português do Brasil. Na etapa de recuperação de informação, foram avaliados cinco modelos de *embeddings* de texto, compreendendo o modelo de código aberto Multilingual E5 e o modelo proprietário da OpenAI1. Para a etapa de geração, foram investigados os modelos de código aberto do Llama e Gemma, bem como os modelos proprietários Sabiá-3 e Sabiazinho-3.

Em uma perspectiva complementar, focada na exploração da estrutura documental, Darcio et al. (2025) apresentaram o LattesRex, um sistema de Q&A, voltado à análise de currículos da Plataforma Lattes, que utiliza a estrutura de metadados para segmentar e recuperar trechos semânticos, em contraste com o *chunking* tradicional. Ao comparar modelos como GPT-4o e a família Gemini 2.5, o estudo evidenciou como a estrutura nativa do documento pode mitigar alucinações e vieses de verbosidade.

Um aspecto importante que une dois desses estudos para o português é a metodologia de avaliação, pois reconhecendo as limitações das métricas automáticas para o português em domínios de alta especificidade, os autores priorizaram a validação humana especializada, que no trabalho de Seabra et al. (2024) baseou-se na avaliação qualitativa realizada por dois especialistas de domínio em 75 contratos. Já no trabalho de Darcio et al. (2025) empregou-se linguistas para avaliar critérios qualitativos como aderência à instrução e qualidade da linguagem.

3 Metodologia

A metodologia adotada é composta por quatro etapas, alinhadas aos componentes do pipeline RAG. Inicialmente, construímos um *dataset* sintético com modelos generativos, utilizado como conjunto mínimo para testar, de maneira controlada, os módulos de recuperação, reordenamento e geração antes da avaliação com especialistas. Em seguida, executamos uma avaliação quantitativa de modelos de *embeddings* e de estratégias de reordenamento, analisando a capacidade de recuperar e priorizar trechos de documentos pertinentes às perguntas do *dataset*. Na terceira etapa, conduzimos uma avaliação com especialistas em cenário de uso, ampliando o *dataset* inicial para uma versão híbrida a partir das interações registradas. Por fim, mantemos o contexto recuperado fixo e comparamos cinco modelos generativos usando BERTScore (F1-Score), isolando o efeito do gerador. A Figura 1 resume o fluxo geral da metodologia.

3.1 Preparação do *Dataset* Sintético

Esta subseção descreve a construção do *dataset* sintético e a preparação do corpus para indexação, que compõem a primeira etapa da metodologia. O objetivo é viabilizar testes controlados dos módulos de recuperação e reordenamento antes da avaliação em cenário real, além de estabelecer uma ligação rastreável entre cada pergunta e sua evidência documental, necessária para as métricas de ranqueamento utilizadas na etapa quantitativa subsequente.

Selecionamos 107 documentos normativos internos provenientes de cinco departamentos hospitalares: Unidade de Saúde Ocupacional e Segurança do Trabalho, Divisão de Recursos Humanos, Serviço de Controle de Infecção Hospitalar, Departamento de Gestão de Ensino e Pesquisa e Setor de Tecnologia da Informação e Saúde Digital (Empresa Brasileira de Serviços Hospitalares - EBSERH, 2025). Esses documentos foram escolhidos por representarem rotinas institucionais frequentes e por apresentarem linguagem normativa (regras, exceções e procedimentos), alinhada ao cenário-alvo do sistema RAG.

A partir desse corpus, geramos automaticamente 90 pares pergunta–resposta com o Gemini 1.5 Flash, que era a versão mais avançada disponível no período de construção do conjunto de dados. Para reduzir a produção de questões excessivamente genéricas e aumentar a cobertura por documento, a geração foi condicionada a trechos de até 1.000 *tokens* extraídos do texto original. Em seguida, realizamos uma revisão manual para remover pares inconsistentes, redundantes ou não ancorados no conteúdo documental. Ao final, 74 pares foram considerados válidos e utilizados como base nas avaliações quantitativas subsequentes.

Após a criação das perguntas, os documentos foram segmentados em unidades menores (*chunks*) para indexação no banco vetorial. Adotamos *chunks* de aproximadamente 100 *tokens*, aplicando

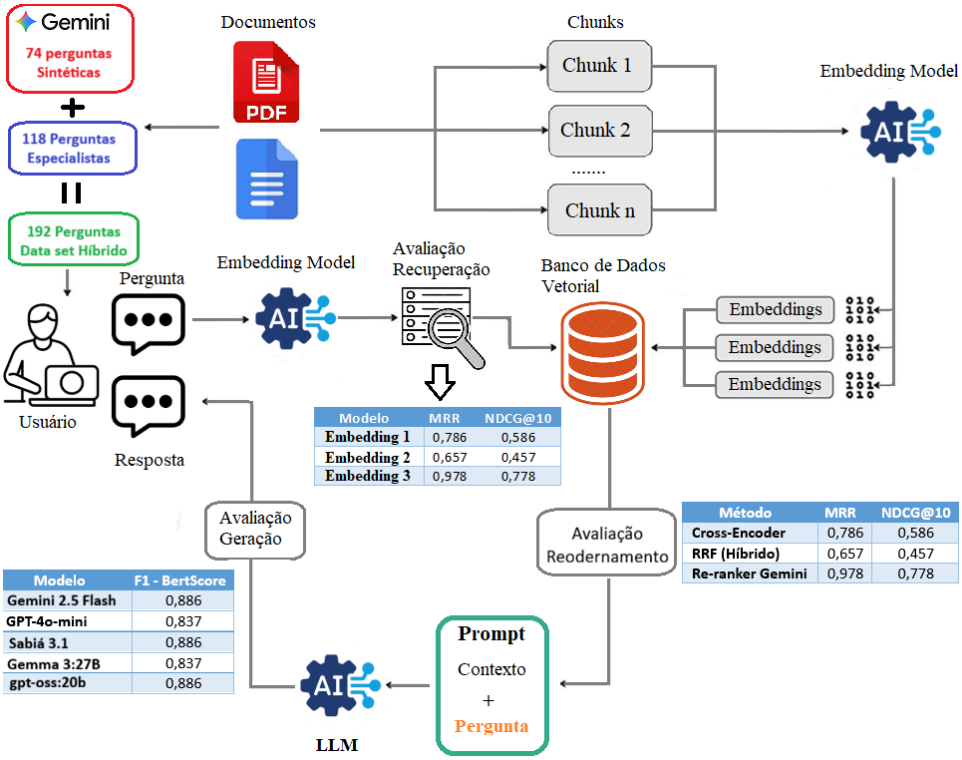

Figura 1: Fluxograma da metodologia. Diagrama da arquitetura do pipeline RAG, destacando o fluxo de dados, as etapas de avaliação e as métricas de desempenho para cada componente.

um limite *suave*: ao atingir esse valor, o corte era deslocado para o próximo ponto final, preservando coesão sintática. Essa configuração mostrou-se mais adequada do que alternativas baseadas em divisão por linha ou por página, além de aproximar a extensão de um parágrafo –frequentemente suficiente para conter a evidência necessária em documentos normativos.

Para aumentar a rastreabilidade, armazenamos também o texto integral da página de origem de cada *chunk* em um campo dedicado no banco de dados. Por fim, identificamos manualmente, para cada uma das 74 perguntas, o *chunk* que contém a evidência correspondente. Assim, cada pergunta foi vinculada ao identificador do *chunk* no índice vetorial, permitindo computar métricas como MRR e NDCG@10 de forma reprodutível na etapa de avaliação quantitativa.

3.2 Avaliação Quantitativa de Modelos de *Embeddings* e Reordenação

Na etapa de recuperação, avaliamos três modelos de *embeddings* para representar tanto as perguntas quanto os *chunks* dos documentos em vetores densos de alta dimensão, permitindo buscas semânticas no banco vetorial. Os modelos analisados foram *all-MiniLM-L6-v2*, *paraphrase-multilingual-MiniLM-L12-v2* e *intfloat/multilingual-e5-small*. Para cada pergunta, calculamos a similaridade de cosseno entre o vetor da pergunta e os vetores dos *chunks* indexados, ranqueando os candidatos por proximidade semântica (Subramanyam and Sangeetha, 2020).

A qualidade do ranqueamento foi medida por duas métricas complementares. A *Mean Reciprocal Rank* (MRR) captura a posição do primeiro *chunk* correto na lista recuperada, enquanto a *Normalized Discounted Cumulative Gain* em $k = 10$ (NDCG@10) avalia a qualidade do topo do ranqueamento, ponderando mais fortemente os resultados nas primeiras posições. Em ambos os casos, consideramos como relevante o *chunk* previamente identificado manualmente como contendo a evidência para a resposta.

O protocolo experimental consistiu em, para cada uma das 74 perguntas do *dataset* sintético, recuperar e ordenar os 100 *chunks* mais similares segundo cada modelo de *embeddings*. Em seguida, registramos a posição do *chunk* relevante (via seu identificador no banco) e calculamos MRR e NDCG@10 agregados sobre o conjunto de perguntas. O uso de 100 candidatos por pergunta tem caráter exploratório: busca-se garantir cobertura suficiente de evidências na lista inicial, permitindo

que a etapa de reordenação opere sobre um conjunto amplo, porém ainda manejável, de candidatos.

Entretanto, a inclusão direta de muitos *chunks* no contexto do modelo gerador implica alto consumo de *tokens* e pode degradar a qualidade da resposta por excesso de informação. Assim, introduzimos uma etapa de reordenação com o objetivo de refinar a priorização dos *chunks* mais relevantes, reduzindo o número de trechos enviados ao LLM e aumentando a probabilidade de incluir a evidência correta no contexto final.

Na comparação de estratégias de reordenação, avaliamos inicialmente os três modelos de *embeddings* na recuperação dos candidatos. Observamos que, após a reordenação, as diferenças dos resultados entre aplicações de diferentes *embeddings* se reduzem substancialmente; por parcimônia experimental, fixamos o modelo com melhor desempenho na etapa de recuperação (intfloat/multilingual-e5-small) e reportamos os resultados de reordenação a partir dessa configuração. As estratégias de reordenação avaliadas foram:

- *Modelo Cross-Encoder*: modelo de relevância que codifica conjuntamente a pergunta e cada *chunk* candidato, produzindo um escore de similaridade mais preciso para reordenar a lista inicial (Puthenputhussery et al., 2025);
- *Reciprocal Rank Fusion* (RRF): método de fusão de rankings que combina listas ordenadas provenientes de diferentes sinais de recuperação (por exemplo, ranqueamento semântico e lexical), resultando em uma ordenação única (Cormack et al., 2009);
- Reordenação baseada em LLM: uso do Gemini 1.5 Flash para reordenar os *chunks* candidatos por meio de um *prompt* estruturado, produzindo uma lista final priorizada (Carraro and Bridge, 2025).

Partindo, para cada pergunta, da lista inicial de 100 *chunks* recuperados na etapa anterior, aplicamos cada estratégia para produzir uma nova ordenação dos candidatos. Em seguida, medimos novamente MRR e NDCG@10 com base na posição do *chunk* relevante (identificado manualmente), a fim de quantificar o ganho de ranqueamento obtido pela reordenação. A Figura 2 apresenta o pipeline dessa etapa.

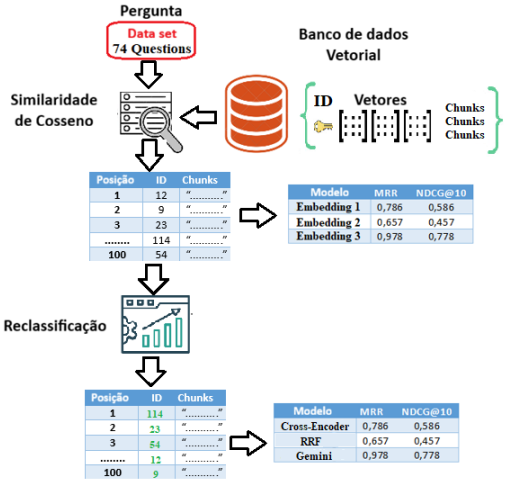

Figura 2: Pipeline da Avaliação da Recuperação.

3.3 Avaliação Qualitativa com Especialistas e Construção do *Dataset* Híbrido

Após a seleção do modelo de *embeddings* e da estratégia de reordenação na etapa quantitativa, configuramos o sistema para validação qualitativa em cenário de uso. Nesta fase, utilizamos o Gemini 2.0 Flash como modelo gerador e submetemos o *chatbot* à avaliação de 12 especialistas do hospital, com o objetivo de verificar sua utilidade percebida e coletar interações reais para ampliar o *dataset* inicial.

Cada especialista formulou perguntas relacionadas à sua rotina e avaliou a resposta como satisfatória ("Sim") ou insatisfatória ("Não"). Quando a resposta era rejeitada, o sistema solicitava a "resposta esperada", permitindo registrar um par pergunta–resposta de referência e apoiar a análise de falhas. Ao final, coletamos 139 interações.

Em termos absolutos, 118 respostas foram classificadas como "Satisfatórias"(Sim), indicando que o sistema foi capaz de recuperar e sintetizar a informação correta na grande maioria dos casos apresentados. Dessa forma, essas 118 perguntas e respostas aprovadas por especialistas foram incorporadas ao *dataset* sintético para compor um *dataset* híbrido, totalizando 192 pares pergunta-resposta utilizados na avaliação comparativa de modelos geradores descrita na próxima subseção. Exemplos representativos de perguntas e respostas obtidas nessa etapa são apresentados na Seção 4.1 (Tabela 1).

3.4 Avaliação Comparativa de Modelos Geradores

O objetivo desta etapa é comparar modelos geradores capazes de produzir respostas em português

no domínio de documentos normativos, de modo a subsidiar futuras iterações do sistema em cenário real. Em particular, buscamos identificar alternativas de LLM que possam ser incorporadas como opções de geração em uma próxima rodada de uso com especialistas, considerando que a escolha do gerador impacta diretamente estilo, completude e adequação linguística das respostas.

Para isolar o efeito do modelo gerador, fixamos o contexto fornecido pelo pipeline de recuperação. Para cada pergunta do *benchmark* (192 pares), selecionamos os 10 *chunks* mais bem ranqueados após a reordenação por RRF e os concatenamos em um único contexto. Esse mesmo contexto, juntamente com um *prompt* de orientação idêntico, foi fornecido a todos os modelos avaliados, garantindo uma comparação controlada. Quando a concatenação excedia o limite de entrada do modelo, aplicamos truncamento por ordem de ranqueamento, preservando os *chunks* mais bem posicionados até o limite do tamanho do contexto do modelo.

Com essa configuração, comparamos cinco modelos de linguagem: Gemini 2.5 Flash, GPT-4o-mini e Sabiá 3.1 (via API), além de Gemma 3:27B e gpt-oss:20b executados localmente. A execução dos modelos locais foi realizada em uma máquina com GPU NVIDIA GeForce RTX 5090 (32 GB VRAM), assegurando memória suficiente para carregamento dos pesos e a execução de inferência eficiente. Por fim, avaliamos quantitativamente a qualidade das respostas por similaridade semântica em relação à resposta de referência, utilizando BERTScore (F1) (Caseli and Nunes, 2024).

4 Resultados e Discussão

Esta seção apresenta os resultados obtidos nos experimentos descritos na metodologia, seguidos de uma análise comparativa e de discussão de suas implicações.

4.1 Análise das Interações com Especialistas

Os registros coletados na validação qualitativa permitem observar, em cenário real, como especialistas formulam consultas em português ao buscar normas e rotinas institucionais, bem como em quais situações o *chatbot* é percebido como útil. A taxa de aceitação observada (118/139) indica que, na maioria dos casos, as respostas foram julgadas adequadas para o objetivo prático do usuário, ainda que se trate de uma medida subjetiva.

Em uma inspeção qualitativa das interações aceitas, notamos predominância de solicitações factuais e procedurais típicas de documentos normativos (e.g., tempos, materiais, condições e exceções), frequentemente expressas em português formal e com estruturas enumerativas. Esse padrão é consistente com o gênero textual do domínio, no qual são comuns formulações deônticas (*deve*, *recomenda-se*), restrições (*exceto*, *quando*) e descrições operacionais em sequência. A Tabela 1 exemplifica esse perfil de consulta e resposta.

Essas observações sugerem que, além de recuperar evidências relevantes, a geração deve preservar características do registro normativo em português, como completude de listas, manutenção de unidades e intervalos e explicitação de condições e exceções.

4.2 Desempenho do Modelo de *Embedding* na Recuperação Inicial

A análise comparativa dos três modelos de *embeddings* revelou uma disparidade substancial na robustez. O modelo intfloat/multilingual-e5-small demonstrou clara superioridade, apresentando uma taxa de falha de 1,4% (falhando em recuperar o fragmento correto em apenas 1 das 74 questões). Em contraste, o modelo paraphrase-multilingual-MiniLM-L12-v2 apresentou uma taxa de falha de 8,1% (6 questões), enquanto o modelo de referência, all-MiniLM-L6-v2 apresentou uma taxa de falha de 9,5% (7 questões).

Essa superioridade foi ainda mais corroborada pela análise quantitativa da métrica MRR, que avalia a posição do primeiro resultado correto. Como mostrado na Tabela 2, o modelo intfloat/multilingual-e5-small alcançou uma MRR média de 0,6662, consideravelmente superior à dos outros dois modelos, ambos em torno de 0,52.

Além da posição do primeiro *chunk* correto como resultado, a qualidade geral da classificação foi avaliada usando a métrica NDCG@10. Os resultados desta análise, também apresentados na Tabela 2, reforçam ainda mais a vantagem do modelo intfloat/multilingual-e5-small, que alcançou uma pontuação média de NDCG@10 de 0,6782, superando tanto o paraphrase-multilingual-MiniLM-L12-v2 com 0,5625 quanto o all-MiniLM-L6-v2 com 0,5445.

Considerando esses resultados, a escolha do modelo intfloat/multilingual-e5-small se justifica não apenas por sua capacidade de minimizar falhas críticas de recuperação, mas também por sua maior eficácia em posicionar documentos relevantes no

Pergunta	Resposta
Qual o tempo médio de duração da higienização simples das mãos com água e sabonete, e qual o tempo médio da fricção antisséptica das mãos com preparações alcoólicas?	A higienização simples das mãos com água e sabonete deve durar em média 40-60 segundos; A fricção antisséptica das mãos dura em média 30 segundos.
Qual o tipo de acomodação para pacientes com tuberculose ativa?	Pacientes com tuberculose ativa devem ser acomodados preferencialmente em quarto privativo, com portas fechadas e saída restrita. A exceção ocorre em casos de coorte, onde pacientes com a mesma patologia podem dividir o quarto, exceto se forem pacientes com suspeita ou confirmação de tuberculose resistente.

Tabela 1: Exemplos representativos de perguntas e respostas obtidas na validação qualitativa, ilustrando consultas típicas do domínio normativo (valores/intervalos, listas e regras com exceções).

topo da lista, conforme demonstrado pelas métricas MRR e NDCG@10. Portanto, esse modelo foi selecionado como base para todas as avaliações subsequentes.

4.3 Análise Comparativa de Métodos de Reordenação

Após definir o modelo de incorporação mais eficaz, a pesquisa concentrou-se na otimização da etapa de reordenação. Nesta fase, os 100 *chunks* mais relevantes recuperados pela busca inicial foram reordenados usando três métodos distintos: *Cross-Encoder*, RRF e reordenação baseada em Gemini 1.5 Flash.

A análise da métrica MRR revelou uma hierarquia de desempenho clara. Como mostrado na Tabela 3, a reordenação baseada no Gemini 1.5 Flash apresentou o maior MRR médio (0,8612) e a maior frequência de acertos perfeitos (MRR = 1), garantindo o primeiro lugar. O RRF também demonstrou um desempenho notavelmente forte, com MRR médio de 0,7857, enquanto o *Cross-Encoder* obteve desempenho inferior (0,6584).

Além disso, a qualidade da classificação foi avaliada usando a métrica NDCG@10. Os resultados, também apresentados na Tabela 3, confirmam a vantagem da reordenação baseada no Gemini 1.5 Flash, que alcançou NDCG@10 médio de 0,8826. O RRF apresentou desempenho semelhante (0,8157), superando novamente o *Cross-Encoder* (0,6906).

A discussão desses resultados nos permite concluir que a reordenação baseada no Gemini 1.5 Flash é a escolha tecnicamente superior para maximizar a precisão da classificação, beneficiando-se das capacidades de um LLM para capturar nuances contextuais. No entanto, o desempenho do RRF é uma das descobertas mais relevantes deste estudo. Por ser um método computacionalmente mais leve e direto, sua capacidade de superar uma abordagem mais sofisticada como o *Cross-Encoder* o posiciona como uma alternativa de baixo custo computacional para aplicações práticas, onde fatores como latência e custo computacional são críticos.

4.4 Avaliação da Qualidade da Geração de Respostas

Nesta fase da avaliação do sistema RAG, foi realizada uma análise comparativa entre os modelos, em que utilizaram-se os 192 pares de perguntas e respostas do *dataset* híbrido, e buscou-se avaliar a qualidade semântica da resposta entregue ao usuário. Assim, para isolar a capacidade de geração de cada modelo, ambos receberam o mesmo contexto, fornecido pelo reordenador RRF.

Além disso, a construção do *Dataset* Híbrido (192 pares) permitiu uma avaliação mais rigorosa dos modelos generativos (Sabiá 3.1, Gemini 2.5, etc). Diferente de *benchmarks* puramente sintéticos, a inclusão de 118 exemplos derivados de interações reais assegura que as métricas de BERTScore reportadas refletem a capacidade dos modelos de responder a questões autênticas e terminologias específicas do cotidiano hospitalar, superando as limitações de datasets gerados artificialmente.

A análise dos resultados com o BERTScore (F1-Score) na Tabela 4 mostra uma clara distinção de desempenho entre os modelos acessados via API e os modelos executados localmente. O modelo Gemini 2.5 Flash obteve melhor desempenho absoluto, com uma média de 0,8185, o qual foi seguido de perto pelo modelo brasileiro Sabiá 3.1 com 0,8085 e pelo GPT-4o-mini com 0,8058.

Já os modelos executados na infraestrutura local apresentaram desempenho inferior, com o gpt-oss:20b atingindo 0,6958 e o Gemma 3:27B registrando 0,6839. Essa diferença sugere que, para este conjunto de perguntas e para este domínio, esses modelos locais podem ter maior dificuldade em sintetizar respostas tão precisas quanto os modelos

	MRR			NDCG@10		
Modelo	**Média**	**Mediana**	**Freq.=1 (%)**	**Média**	**Mediana**	**Freq.=1 (%)**
all-MiniLM-L6-v2	0,5225	0,5	40,5405	0,5445	0,6309	36,4865
paraphrase-multilingual-MiniLM-L12-v2	0,5200	0,5	40,5405	0,5625	0,6220	37,8378
intfloat/multilingual-e5-small	**0,6662**	**1,0**	**58,1081**	**0,6782**	**1,0000**	**52,7027**

Tabela 2: Estatísticas de ranqueamento por modelo de *embedding* na recuperação inicial (MRR e NDCG@10).

	MRR			NDCG@10		
Método	**Média**	**Mediana**	**Freq.=1 (%)**	**Média**	**Mediana**	**Freq.=1 (%)**
Cross-Encoder	0,6584	1,0	54,0541	0,6906	0,9599	50,0000
RRF	0,7857	1,0	67,5676	0,8157	1,0000	64,8649
Gemini 1.5 Flash	**0,8612**	**1,0**	**78,3784**	**0,8826**	**1,0000**	**75,6757**

Tabela 3: Estatísticas de ranqueamento por método de reordenação (MRR e NDCG@10).

de maior desempenho avaliados, possivelmente por diferenças de capacidade, alinhamento e adequação linguística ao português do domínio normativo, e não pelo modo de execução em si. Assim, para a tarefa de geração de respostas neste projeto, o Gemini 2.5 Flash mostrou-se ligeiramente superior, estabelecendo-se como a opção de melhor desempenho entre os modelos avaliados.

Modelo	**Média (BERTScore (F1-Score))**
Gemini 2.5 Flash	**0,8185**
Sabiá 3.1	**0,8085**
GPT-4o-mini	**0,8058**
gpt-oss:20b	**0,6958**
Gemma 3:27B	**0,6839**

Tabela 4: Estatísticas da métrica BERTScore (F1-Score) para modelos de geração.

5 Conclusões

A avaliação por etapas do pipeline RAG evidenciou a utilidade desta abordagem metodológica para otimizar o equilíbrio entre a recuperação e geração de respostas do sistema. Os testes realizados nas três etapas distintas – recuperação de informação, reordenamento de documentos e geração de resposta – demonstraram que a análise granular de cada componente foi fundamental para identificar a combinação de modelos com melhor custo-benefício.

Dessa forma, conclui-se que o modelo de *embedding* intfloat/multilingual-e5-small, dentre os modelos avaliados, é a escolha mais promissora para a etapa de recuperação inicial, minimizando falhas de recuperação e melhorando o ranqueamento dos *chunks* relevantes. Na etapa de reordenamento, embora a reordenação baseada em Gemini 1.5 Flash ofereça a maior precisão, o RRF emerge como uma alternativa de custo-benefício superior, pois sua eficiência se mantém com desempenho competitivo e menor complexidade computacional.

Finalmente, na etapa de geração, observou-se que os modelos de maior desempenho apresentaram resultados ligeiramente superiores na métrica BERTScore (F1-Score), com destaque para o Gemini 2.5 Flash, que se estabelece como a opção de melhor desempenho e consistência entre os modelos avaliados para este projeto. Portanto, a escolha da arquitetura final prioriza a eficiência, combinando intfloat/multilingual-e5-small com a reordenação com RRF e o gerador Gemini 2.5 Flash para um sistema robusto, preciso e computacionalmente viável.

Para trabalhos futuros, diversas possibilidades de pesquisa se apresentam para expandir e aprimorar o sistema. Pretende-se ampliar a avaliação qualitativa do *chatbot* com profissionais de mais setores do hospital e com maior número de participantes, permitindo coletar *feedback* adicional sobre usabilidade da interface, relevância contextual das respostas e satisfação geral dos usuários, fortalecendo a validação em cenário de uso real.

Agradecimentos

Este trabalho foi apoiado pelo Instituto Federal do Rio Grande do Sul (IFRS), Empresa Brasileira e Serviços Hospitalares (EBSERH) e Hospital Escola da UFPEL. Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código Financeiro 001. Gostaríamos de agradecer à FAPERGS -

Brasil pelo Apoio Financeiro, Contrato de Outorga 22/2551-0000598-5. Reconhecemos com gratidão o apoio da NVIDIA Corporation com a doação da GPU Titan X Pascal usada para esta pesquisa.

References

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

I. Alonso, M. Oronoz, and R. Agerri. 2024. MedexpQA: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

M. Alshammary, M. N. Uddin, and L. Khan. 2024. Rfpg: Question-answering from low-resource language (arabic) texts using factually aware rag. In *Proceedings of the 2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC 2024)*, pages 107–116. IEEE.

Orlando Ayala and Patrice Bechard. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238.

Diego Carraro and Derek Bridge. 2025. Enhancing recommendation diversity by re-ranking with large language models. *ACM Transactions on Recommender Systems*, 4(2):1–40.

H.M. Caseli and M.G.V. Nunes, editors. 2024. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edition. BPLN. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao.

O. Cederlund, S. Alawadi, and F. M. Awaysheh. 2024. Llmrag: An optimized digital support service using llm and retrieval-augmented generation. In *Proceedings of the 9th International Conference on Fog and Mobile Edge Computing (FMEC 2024)*, pages 54–62, Malmö, Sweden.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. *Preprint*, arXiv:2407.01463.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Murilo Vargas da Cunha, Marilia Rosa Silveira, Brenda Salenave Santana, Larissa Astrogildo Freitas, and Ulisses Brisolara Corrêa. 2025. Optimizing and evaluating a retrieval-augmented generation system for normative document retrieval in hospital settings. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 385–393. SBC.

Lucas Darcio, Karina Santos, Amanda Spellen, Esther Soares, Livy Real, and Altigran Silva. 2025. Lattesrex: Building chatbots for semi-structured documents. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 125–136, Porto Alegre, RS, Brasil. SBC.

Empresa Brasileira de Serviços Hospitalares - EBSERH. 2025. Estrutura administrativa — hu-ufsc. https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sul/hu-ufsc/governanca/estrutura-administrativa. Acesso em: 13 jul. 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):38.

M. Maryamah, M. M. Irfani, E. B. Tri Raharjo, N. A. Rahmi, M. Ghani, and I. K. Raharjana. 2024. Chatbots in academia: A retrieval-augmented generation approach for improved efficient information access. In *Proceedings of the 16th International Conference on Knowledge and Smart Technology (KST 2024)*, pages 259–264, Krabi, Thailand.

Luiz Sabiano Ferreira Medeiros and Hilário Tomaz Alves de Oliveira. 2025. Comparação de modelos de embeddings e llms para geração aumentada por recuperação em português. In *Seminário Integrado de Software e Hardware (SEMISH)*, pages 429–440. SBC.

R. Nai, E. Sulis, I. Fatima, and R. Meo. 2024. Large language models and recommendation systems: A proof-of-concept study on public procurements. In *Proc. of the 29th Int. Conf. on Applications of Natural Language to Information Systems (NLDB 2024), Part II*, pages 280–290, Turin, Italy.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models. *Preprint*, arXiv:2307.06435.

S. Obaid and N. Z. Bawany. 2024. Seerahgpt: Retrieval augmented generation based large language model. In *Proceedings of the 18th International Conference on Open Source Systems and Technologies (ICOSST 2024)*, pages 1–7, Lahore, Pakistan.

S. S. T. Oliveira, D. Fazzioni, and D. O. C. Ferreira. 2025. Grandes modelos de linguagem. In *In: Kudo, T. N. et al.* Cegraf UFG, Goiânia. E-book (254 p.). ISBN 978-85-495-1096-9.

Shanli Ouyang. 2025. Core applications and techniques of rag for low-resource languages. *Science and Technology of Engineering, Chemistry and Environmental Protection*, 1(1).

Ajit Puthenputhussery, Changsung Kang, Alessandro Magnani, Tian Zhang, Hongwei Shang, Nitin Yadav, Prijith Chandran, Bhavin Madhani, Yuan-Tai Fu, He Wang, and 1 others. 2025. Large scale deployment of bert based cross encoder model for reranking in walmart search engine. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4365–4369.

S. Raschka. 2024. *Build a Large Language Model (From Scratch)*. From Scratch. Manning.

Mubashar Raza, Zarmina Jahangir, Muhammad Bilal Riaz, Muhammad Jasim Saeed, and Muhammad Awais Sattar. 2025. Industrial applications of large language models. *Scientific Reports*, 15:13755.

A. Seabra, C. Cavalcante, J. Nepomuceno, L. Lago, N. Ruberg, and S. Lifschitz. 2024. Contrato360 2.0: A document and database-driven question-answer system using large language models and agents. *arXiv preprint arXiv:2412.17942*.

K. S. K. Subramanyam and S. Sangeetha. 2020. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of Biomedical Informatics*, 101:103323.

O. C. Wijaya and A. Purwarianti. 2024. An interactive question-answering system using large language model and retrieval-augmented generation in an intelligent tutoring system on the programming domain. In *Proceedings of the 2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–6, Singapore.

H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu. 2025. Evaluation of retrieval-augmented generation: A survey. In *Proceedings of the Big Data*, pages 102–120, Singapore. Springer.

Z. Zhu, Y. Wang, H. Liu, X. Chen, and M. Zhang. 2024. Enhancing large language models with knowledge graphs for robust question answering. In *Proceedings of the 30th IEEE International Conference on Parallel and Distributed Systems (ICPADS 2024)*, pages 262–269, Belgrade, Serbia. IEEE.