APS-RAG-Benchmark: Um Recurso Inicial para Avaliação de RAG na Atenção Primária à Saúde

Claudio M. V. de Andrade[1], **Gestefane Rabbi Magalhães**[1], **Raiane Asevedo**[1]
Julia Paes[1], **Isaias José Ramos Oliveira**[1], **Adriana Pagano**[1]
Zilma Reis[1], **Marcos André Gonçalves**[1]

[1]Unversidade Federal de Minas Gerais (UFMG)

{claudio.valiense,gestefane,juliapaes,mgoncalv}@dcc.ufmg.br

{apagano,raiasevedo,zilma}@ufmg.br, isaias@medicina.ufmg.br

Resumo. *Este trabalho apresenta o APS-RAG-Benchmark, um benchmark preliminar para avaliação de sistemas Retrieval-Augmented Generation (RAG) no domínio do e-SUS Atenção Primária à Saúde (APS). O benchmark reúne um dataset validado por especialistas, composto por perguntas reais de tickets de suporte, fontes documentais relevantes e respostas geradas por modelos de linguagem e revisadas por humanos. Além de disponibilizar um recurso para um domínio ainda pouco explorado em português, o trabalho propõe um protocolo experimental que avalia separadamente as etapas de recuperação e geração. Os resultados mostram que a combinação de estratégias léxicas (BM25) e semânticas melhora significativamente a recuperação, elevando o MRR de 0,298 para 0,596. Além disso, observou-se alinhamento frequente entre as fontes recuperadas e as consideradas corretas pelos especialistas, com até 69% das respostas avaliadas como adequadas.*

Abstract. *This work presents APS-RAG-Benchmark, a preliminary benchmark for evaluating Retrieval-Augmented Generation (RAG) systems in the domain of e-SUS Primary Health Care (APS). The benchmark compiles a dataset validated by experts, consisting of real questions from support tickets, relevant document sources, and answers generated by large language models and reviewed by human specialists. In addition to providing a resource for a domain that remains underexplored in Portuguese, the work proposes an experimental protocol that evaluates the retrieval and generation stages separately. The results show that combining lexical (BM25) and semantic strategies significantly improves retrieval performance, increasing MRR from 0.298 to 0.596. Furthermore, frequent alignment was observed between retrieved sources and those considered relevant by experts, with up to 69% of the generated answers being rated as adequate.*

1. Introdução

Solicitações de suporte técnico, comumente denominadas *tickets*, são geradas a partir de dúvidas e problemas relatados por usuários dos softwares do e-SUS Atenção Primária à Saúde (APS)[1]. Nesse contexto, sistemas automáticos de resposta baseados em grandes modelos de linguagem (LLMs) surgem como uma alternativa promissora para apoiar atendentes humanos e reduzir o custo operacional do suporte técnico.

[1]https://sisaps.saude.gov.br/sistemas/esusaps/

Recentemente, arquiteturas baseadas em *Retrieval-Augmented Generation* (RAG) [Lewis et al. 2020] tornaram-se predominantes em aplicações de atendimento automatizado [Gao et al. 2023]. Nesses sistemas, a qualidade das respostas depende diretamente da eficácia da recuperação de documentos relevantes e da capacidade do modelo gerador em sintetizar essas informações [Ram et al. 2023]. Apesar dos avanços recentes, a avaliação de sistemas RAG ainda permanece desafiadora em domínios especializados e em cenários com baixa disponibilidade de recursos, como o português, nos quais *datasets* públicos e protocolos experimentais são escassos.

No domínio do e-SUS APS, esse desafio é particularmente relevante devido à natureza especializada das dúvidas e à ausência de recursos padronizados para treinamento e avaliação. Além disso, perguntas reais frequentemente apresentam ruído, múltiplas intenções e descrições incompletas, dificultando tanto a recuperação de fontes relevantes quanto a geração de respostas adequadas.

Neste contexto, este trabalho apresenta o **APS-RAG-Benchmark**, um recurso preliminar para avaliação de sistemas RAG no domínio do e-SUS APS. Adotando uma perspectiva *data-centric* [Zha et al. 2025], construímos um *dataset* a partir de perguntas reais extraídas do sistema de atendimento, combinadas com respostas geradas por LLMs e posteriormente revisadas por especialistas humanos. O recurso associa perguntas, respostas e fontes documentais relevantes, permitindo analisar separadamente as etapas de recuperação e geração[2].

Também propomos um protocolo experimental inicial combinando métricas automáticas e avaliação qualitativa por especialistas do domínio. Com base nesse protocolo, realizamos uma análise exploratória comparando estratégias de recuperação léxica, semântica e fusão de rankings. Diante desse contexto, investigamos as seguintes questões de pesquisa:

- **RQ1:** Qual é o nível de eficácia do sistema de recuperação para identificar fontes relevantes à construção de respostas no domínio do e-SUS APS?
- **RQ2:** Qual é a qualidade das respostas geradas segundo critérios de correção, completude, clareza, efetividade e utilidade?

Os resultados indicam que a combinação de estratégias léxicas e semânticas melhora significativamente a recuperação de documentos relevantes, enquanto a avaliação qualitativa mostra que uma parcela substancial das respostas foi considerada adequada pelos especialistas. De forma geral, os achados sugerem que recursos estruturados e protocolos de avaliação representam um passo promissor para o estudo de sistemas RAG em domínios especializados em português.

Este estudo concentra-se principalmente na análise da etapa de recuperação e em seu impacto sobre a geração, deixando comparações mais amplas entre arquiteturas generativas para trabalhos futuros.

2. Fundamentação e Trabalhos Relacionados

Em [Andrade et al. 2025], os autores investigam o uso de dados sintéticos, incluindo LLMs, para geração de paráfrases com o objetivo de mitigar o desbalanceamento entre

[2]Disponibilizaremos publicamente o código-fonte, o *dataset*, hiperparâmetros, casos de acertos e erros e os *prompts* utilizados em `https://github.com/claudiovaliense/e-sus-aps-chat`

classes em tarefas de classificação de textos em português a partir de *tickets* de atendimento. Essa linha de trabalho foca na melhoria de modelos preditivos por meio de aumento de dados. Em contraste, o presente trabalho adota uma perspectiva *data-centric* voltada à construção de um *dataset* validado por especialistas para avaliação de sistemas de recuperação e geração. Propomos uma metodologia em que respostas são geradas por um LLM e posteriormente avaliadas e revisadas por especialistas, permitindo análises controladas dos componentes do *pipeline*. Complementarmente, [Lebeis et al. 2025] explora o uso de *prompts* com LLMs para gerar mensagens personalizadas no contexto do e-SUS APS, reforçando a relevância de avaliações estruturadas como as propostas neste trabalho.

Sistemas baseados em RAG têm sido amplamente empregados em tarefas que demandam acesso a conhecimento externo [Gao et al. 2023], nas quais a qualidade das respostas depende diretamente da eficácia da recuperação. Métodos léxicos, como BM25 [Robertson & Zaragoza 2009, Fernandes et al. 2007], são eficazes na correspondência exata de termos, enquanto abordagens semânticas baseadas em *embeddings*, como HNSW [Malkov & Yashunin 2019], permitem capturar relações conceituais entre consultas e documentos. Estudos como [Thakur et al. 2021] demonstram que a combinação desses sinais melhora significativamente a recuperação. Nesse contexto, técnicas de fusão como ZMUV e CombMNZ [Montague & Aslam 2002, França et al. 2025] aumentam a robustez do ranqueamento ao favorecer documentos recuperados simultaneamente por múltiplos métodos, sendo utilizadas neste trabalho para analisar seu impacto na qualidade das respostas geradas.

3. Metodologia Proposta

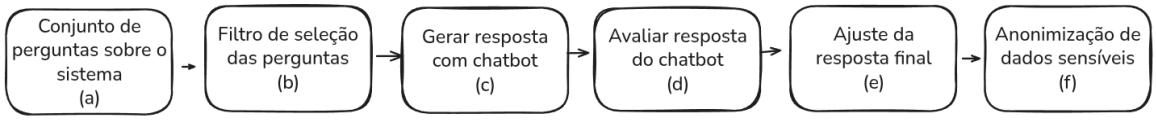

Figura 1. Fluxo geral de construção do dataset e protocolo de avaliação.

A Figura 1 apresenta o fluxo utilizado para construção do APS-RAG-Benchmark e do protocolo experimental adotado neste trabalho. Inicialmente, foi coletado um conjunto de perguntas reais a partir do sistema de atendimento de *tickets* do e-SUS APS (Figura 1(a)). Em seguida, realizou-se uma etapa de filtragem para selecionar apenas perguntas compatíveis com o escopo de um sistema automático de resposta (Figura 1(b)). As perguntas selecionadas foram então submetidas a um sistema baseado em RAG para geração das respostas iniciais (Figura 1(c)). Posteriormente, especialistas do domínio avaliaram a relevância das fontes recuperadas e a qualidade das respostas geradas, podendo revisar ou ajustar as respostas finais (Figura 1(d-e)). Por fim, foi realizada uma etapa de anonimização manual de dados sensíveis (Figura 1(f)). Cabe destacar que apenas a etapa de geração das respostas foi realizada pelo modelo de linguagem, enquanto todas as demais etapas foram conduzidas por avaliadores humanos.

3.1. Seleção das perguntas

O sistema de gerenciamento de *tickets* registra dúvidas relacionadas ao uso do e-SUS APS, posteriormente tratadas por atendentes especialistas. Devido à heterogeneidade das solicitações, muitas perguntas não eram adequadas para um sistema automático de

resposta, seja por exigirem intervenção humana específica ou por não possuírem suporte suficiente na documentação oficial. Dessa forma, realizamos uma etapa de filtragem manual conduzida por dois especialistas do suporte técnico do e-SUS APS. Do total de 119 tickets analisados, foram mantidas 49 perguntas cuja resolução pudesse ser suportada pela documentação oficial, excluindo solicitações dependentes de intervenção administrativa ou contexto específico do usuário.

3.2. Pipeline de recuperação e geração

O sistema utilizado neste estudo adota uma arquitetura RAG composta por recuperação léxica (BM25) [Robertson & Zaragoza 2009] e recuperação semântica baseada em *embeddings* com HNSW [Malkov & Yashunin 2019], ambos retornam os top-k candidatos ($k = 5$). A combinação entre recuperação léxica e semântica foi adotada devido à natureza heterogênea das consultas do e-SUS APS, que frequentemente combinam termos técnicos específicos e descrições textuais mais livres. Os resultados recuperados são posteriormente combinados por fusão de *rankings* utilizando ZMUV e CombMNZ [**?**].

A partir dos documentos mais relevantes, constrói-se o contexto fornecido ao modelo gerador (LLaMA 3.1 8B, temperatura=0, demais parâmetros com valor padrão), escolhido por representar uma configuração aberta amplamente utilizada em aplicações locais de RAG com custo computacional reduzido. Para a recuperação semântica, utilizou-se o modelo de embedding Qwen3-Embedding-0.6B (dimensão vetorial de 1024 posições), selecionado por seu desempenho em tarefas de similaridade semântica em português. A separação entre modelo de embedding e modelo gerador é intencional neste estudo, permitindo avaliar a etapa de recuperação de forma independente da geração, conforme proposto pelo protocolo experimental.

3.3. Avaliação qualitativa

Cada par pergunta–resposta foi avaliado por três especialistas independentes da equipe de suporte do e-SUS APS, com experiência prática no atendimento de dúvidas relacionadas ao sistema e à documentação oficial utilizada no suporte, sendo a consolidação final realizada por consenso. A avaliação considerou tanto a relevância das fontes recuperadas quanto critérios qualitativos relacionados à correção, completude, efetividade, clareza e utilidade das respostas geradas. Além da avaliação, os especialistas puderam revisar e ajustar manualmente as respostas produzidas pelo sistema.

3.4. Anonimização dos dados

Após a etapa de avaliação, realizamos uma inspeção manual para remoção de informações sensíveis, como nomes e CPF, substituídos pela *tag* "[[ANONIMIZADO]]". As perguntas, respostas e avaliações compõem a versão final do dataset utilizada nos experimentos.

4. Resultados Experimentais

Após a aplicação do protocolo descrito na Seção 3, obtivemos 49 pares de perguntas e respostas validados por especialistas, compondo a versão inicial do APS-RAG-Benchmark. Embora o número de instâncias ainda seja reduzido, a construção do dataset envolveu avaliação manual especializada das fontes recuperadas e das respostas geradas, realizada por três avaliadores independentes com consolidação por consenso, permitindo uma análise qualitativa inicial dos componentes de recuperação e geração.

4.1. RQ1: Eficácia da recuperação

Das 49 perguntas avaliadas, 6 não apresentaram fontes relevantes para o modelo, resultando em 43 perguntas com suporte documental. Observou-se que a maioria das respostas foi construída a partir de poucas fontes documentais, indicando que o sistema opera com contextos relativamente compactos. Em particular, 14 perguntas utilizaram 2 fontes para construção das respostas, 10 utilizaram 3 fontes, 8 utilizaram apenas 1 fonte, enquanto apenas 5 perguntas demandaram 5 fontes documentais.

Além disso, verificou-se alinhamento frequente entre as primeiras posições do *ranking* e as fontes consideradas relevantes pelos avaliadores. A primeira posição (fonte 1), associada ao maior escore do modelo, foi julgada relevante em 23 perguntas, seguida por 16, 8, 5 e 2 ocorrências nas posições subsequentes. Esses resultados indicam que a recuperação constitui um componente central para o desempenho do pipeline RAG no domínio do e-SUS APS, sugerindo que documentos posicionados nas primeiras posições do ranking tendem a concentrar as principais evidências utilizadas na geração das respostas.

4.2. RQ2: Qualidade das respostas geradas

Para avaliar a qualidade das respostas geradas, utilizamos critérios qualitativos relacionados à correção, completude, efetividade, clareza e utilidade. A Tabela 1 resume os percentuais das avaliações positivas, o grau de concordância de 2 ou mais avaliadores e o valor Gwet's AC1 [Gwet 2008] observados entre os três avaliadores especialistas. Optamos pelo uso de Gwet's AC1 devido às características da tarefa, que envolve três avaliadores, três classes de avaliação e critérios qualitativos subjetivos potencialmente desbalanceados. Embora, assim como o Fleiss's Kappa, o Gwet's AC1 também seja uma métrica baseada em concordância corrigida pelo acaso, ele tende a apresentar maior estabilidade em cenários com distribuições assimétricas entre categorias e elevada concentração de respostas em classes dominantes, reduzindo o chamado paradoxo do Kappa [Zec et al. 2017, Wongpakaran et al. 2013]. Em comparação ao Krippendorff's Alpha [Krippendorff 2011], o Gwet's AC1 é mais estável em conjuntos reduzidos e com múltiplos avaliadores. Os resultados indicam concordância moderada, compatível com a subjetividade inerente à avaliação qualitativa das respostas.

Tabela 1. Percentual da avaliação positiva, concordância entre dois ou mais avaliadores e valor de Gwet's AC1

Critério	Porcentagem da avaliação positiva	Concordância entre dois ou mais avaliadores	Gwet's AC1
Correção	60.5%	87.8%	0.35
Completude	55.8%	89.8%	0.44
Efetividade	55.8%	89.8%	0.38
Clareza	69.8%	89.8%	0.27
Utilidade	55.8%	81.6%	0.34

Os resultados mostram predominância de avaliações positivas em todos os critérios, com destaque para clareza. Apesar disso, observam-se limitações em cenários que exigem maior precisão interpretativa, indicando oportunidades para aprimoramento tanto da recuperação quanto da geração.

4.3. Estudo de ablação

Além da análise qualitativa, avaliamos a etapa de recuperação por meio do *Mean Reciprocal Rank* (MRR). A Tabela 2 compara três estratégias: busca léxica (BM25), busca semântica baseada em *embeddings* e fusão de rankings.

Tabela 2. Resultados do estudo de ablação.

Estratégia de Recuperação	MRR
Léxica (BM25)	0.298
Semântica (Embedding + KNN)	0.356
Combinada (Fusão de Rankings)	0.596

Os resultados evidenciam a complementaridade entre sinais léxicos e semânticos. Enquanto o BM25 é eficaz na correspondência de termos técnicos, a recuperação semântica captura relações conceituais e variações terminológicas. A combinação das abordagens apresentou o melhor desempenho, indicando que estratégias híbridas são particularmente relevantes em domínios especializados e com consultas potencialmente ruidosas.

4.4. Análise de erros

A análise qualitativa revelou quatro categorias principais de falhas: (i) interpretação incorreta de siglas e termos técnicos; (ii) desalinhamento entre a intenção da pergunta e o conteúdo recuperado; (iii) limitações na segmentação dos documentos (*chunking*); e (iv) perguntas contendo múltiplas intenções em uma única interação. De forma geral, os erros observados reforçam que limitações na recuperação impactam diretamente a qualidade das respostas geradas, especialmente em cenários reais envolvendo consultas ambíguas ou documentação altamente especializada.

5. Conclusões e Trabalhos Futuros

Este trabalho apresentou um recurso para avaliação de sistemas baseados em RAG no domínio do e-SUS APS. Como principal contribuição, propomos o APS-RAG-Benchmark, abrangendo um *dataset* preliminar construído a partir de perguntas reais e respostas revisadas por especialistas e um protocolo experimental voltado à análise separada das etapas de recuperação e geração. Os resultados indicam que a qualidade das respostas depende fortemente da eficácia da recuperação, especialmente em cenários envolvendo consultas ruidosas, múltiplas intenções e documentação especializada. Estratégias híbridas de recuperação, combinando sinais léxicos e semânticos, apresentaram os melhores resultados.

De forma geral, os achados sugerem que recursos estruturados e protocolos de avaliação representam um passo promissor para análises mais sistemáticas de sistemas RAG em português e em domínios especializados com baixa disponibilidade de dados públicos. Como limitações, destacam-se o tamanho ainda reduzido do *dataset* e o foco em um único domínio especializado. Os resultados devem ser interpretados como uma análise exploratória inicial voltada à construção de recursos e protocolos de avaliação para sistemas RAG no domínio do e-SUS APS. Como trabalhos futuros, pretendemos ampliar o *dataset*, investigar estratégias mais robustas de recuperação e geração.

Uso de IA Generativa

LLMs foram utilizadas no pipeline experimental e para apoio à melhoria textual do manuscrito.

Financiamento e Agradecimentos

Ministério da Saúde, Secretaria de Atenção Primária à Saúde (TED 31/2022, Processo 25000.058753/2022-59), CNPq, FAPEMIG e INCT-TILD-IAR (grant # 408490/2024-1).

Referências

Andrade, C. M., Magalhães, G., Asevedo, R., Paes, J., Oliveira, I., Pagano, A., Reis, Z., & Gonçalves, M. A. (2025). Estudo do impacto de dados sintéticos e paráfrases na mitigação do desbalanceamento em tarefas de classificação de textos em português com baixa amostragem. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*.

Fernandes, D., de Moura, E. S., Ribeiro-Neto, B., da Silva, A. S., & Gonçalves, M. A. (2007). Computing block importance for searching on web sites. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*.

França, C., Rabbi, G., Salles, T., Cunha, W., Rocha, L., & Gonçalves, M. A. (2025). Ranking-based fusion algorithms for extreme multi-label text classification (xmtc).

Gao, Y. et al. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability. Technical report, Annenberg School for Communication, University of Pennsylvania.

Lebeis, G., Reis, Z., Oliveira, I., & Pereira, F. (2025). Engagement strategies for digital literacy to support the brazilian e-sus primary care system. In *MEDINFO 2025*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Malkov, Y. A. & Yashunin, D. A. (2019). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Montague, M. H. & Aslam, J. A. (2002). Evaluating score normalization methods in data fusion. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ram, O. et al. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*.

Robertson, S. & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gupta, A. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*.

Zec, S., Soriani, N., Comoretto, R. I., & Baldi, I. (2017). The paradox of cohen's kappa. *Open Nursing Journal*, 11:211–218.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric artificial intelligence: A survey. *ACM Comput. Surv.*