

Recuperando Perguntas em vez de Trechos: Uma Abordagem Question-Augmented RAG para Perguntas sobre Medicamentos

Claudio M.V. de Andrade, Raiane Asevedo, Isabella Schettini, Zilma Reis, Marcos A. Gonçalves
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil

claudio.valiense@dcc.ufmg.br,raiaasevedo@ufmg.br,silvaivg@ufmg.br,zilma@ufmg.br,mgoncalv@dcc.ufmg.br

ABSTRACT

Large Language Models (LLMs) have been increasingly employed for healthcare question answering, particularly in applications requiring access to specialized knowledge. This paper investigates the impact of replacing traditional chunk-based indexing with synthetic question indexing on retrieval quality in Retrieval-Augmented Generation (RAG). Using Portuguese drug package inserts from the Brazilian Health Regulatory Agency (Anvisa), we compare three settings: zero-shot inference, conventional RAG, and Question-Augmented Retrieval-Augmented Generation (QA-RAG), in which retrieval is performed over synthetic questions generated from document chunks. The evaluation comprises 90 automatically generated questions covering the 30 most popular medications in Brazil, with responses assessed by both healthcare specialists and an LLM-as-a-Judge. Results show that context retrieval substantially improves answer quality over zero-shot inference. Moreover, QA-RAG consistently outperforms conventional RAG, achieving a 44% improvement in the human evaluation while producing higher semantic retrieval scores. Although question-based indexing introduces additional computational cost, the results indicate that it provides a favorable trade-off between retrieval effectiveness and computational overhead for pharmaceutical question answering in Portuguese.

KEYWORDS

Retrieval-Augmented Generation, Question Answering, Pharmaceutical Information, LLM-as-a-Judge.

1 INTRODUÇÃO

Sistemas baseados em *Retrieval-Augmented Generation* (RAG) têm se consolidado como uma das principais estratégias para reduzir alucinações e ampliar a confiabilidade de Modelos de Linguagem de Grande Porte (*Large Language Models* – LLMs), especialmente em domínios que dependem de conhecimento especializado, como a área da saúde [11]. Ao complementar o conhecimento paramétrico das LLMs com informações recuperadas dinamicamente de bases externas, essas abordagens permitem produzir respostas fundamentadas em documentos de referência atualizados.

No domínio farmacêutico, bulas de medicamentos constituem uma importante fonte oficial de informação clínica, reunindo dados sobre indicações, contraindicações, posologia, interações medicamentosas e efeitos adversos. Apesar dos avanços recentes em aplicações de RAG para a saúde, ainda existem poucos estudos voltados especificamente para sistemas de perguntas e respostas sobre medicamentos em português e, principalmente, para a investigação de como diferentes estratégias de recuperação influenciam a qualidade das respostas produzidas.

A maioria das abordagens RAG indexa diretamente segmentos (*chunks*) dos documentos. Recentemente, entretanto, trabalhos têm proposto substituir essa representação por perguntas sintéticas geradas automaticamente por LLMs [10]. Nessa abordagem, denominada neste trabalho de *Question-Augmented Retrieval-Augmented Generation* (QA-RAG), a consulta do usuário é comparada semanticamente com perguntas previamente associadas aos documentos, em vez dos próprios *chunks*. Embora essa estratégia apresente resultados promissores, permanece pouco compreendido em que medida a substituição da unidade de indexação melhora efetivamente a recuperação semântica, a qualidade das respostas geradas e o custo computacional do sistema.

Neste trabalho investigamos essa questão utilizando bulas da Agência Nacional de Vigilância Sanitária (Anvisa) como fonte de conhecimento. Comparamos três cenários: (i) *Zero-shot*, sem recuperação de contexto; (ii) RAG convencional, baseado na recuperação de *chunks*; e (iii) QA-RAG, baseado na recuperação de perguntas sintéticas. Os experimentos utilizam 90 perguntas sobre os 30 medicamentos mais populares no Brasil. As respostas são produzidas pela mesma LLM e avaliadas tanto por especialistas da área da saúde quanto por uma abordagem baseada em *LLM-as-a-Judge* (abordagem que utiliza o LLM como avaliador). Também analisamos o custo computacional das diferentes etapas do pipeline e disponibilizaremos publicamente o código-fonte, o conjunto de dados, a documentação e os *prompts* utilizados.¹

Diante desse contexto, investigamos as seguintes questões de pesquisa:

- **RQ1:** Qual é o impacto da recuperação de contexto na qualidade das respostas geradas em comparação a uma abordagem *Zero-shot*?
- **RQ2:** Qual é o custo computacional associado à substituição da indexação baseada em *chunks* por perguntas sintéticas?

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

¹<https://github.com/claudiovaliense/medicamentos>

- **RQ3:** Em que medida substituir a indexação baseada em *chunks* por perguntas sintéticas melhora a recuperação semântica e a qualidade das respostas, considerando avaliações realizadas por especialistas humanos e por *LLM-as-a-Judge*?

Os resultados mostram que a recuperação de contexto melhora significativamente a qualidade das respostas em relação ao cenário *Zero-shot*. Mais importante, observamos que a substituição da indexação baseada em *chunks* por perguntas sintéticas produz ganhos consistentes tanto na recuperação semântica quanto na qualidade das respostas, segundo especialistas humanos e *LLM-as-a-Judge*. Esses resultados sugerem que a escolha da unidade de indexação constitui um fator determinante para sistemas RAG em domínios especializados, indicando que representar explicitamente as possíveis necessidades informacionais dos usuários pode ser mais eficaz do que indexar diretamente segmentos documentais, ainda que essa estratégia implique maior custo durante a construção do índice.

2 TRABALHOS RELACIONADOS

Sistemas de *Retrieval-Augmented Generation* (RAG) complementam o conhecimento paramétrico de LLMs por meio da recuperação de documentos externos, reduzindo alucinações e melhorando a precisão factual das respostas [9]. Representações e unidades de conteúdo alternativas têm sido exploradas para organizar e recuperar informação textual [7, 17–19]. Mais recentemente, diversos trabalhos têm investigado o uso de perguntas sintéticas ou técnicas de expansão de consultas geradas por LLMs para substituir ou complementar a indexação baseada em *chunks* [8, 10, 13, 20]. Essas abordagens aproximam a recuperação da forma como usuários formulam suas consultas, mas ainda são escassos estudos que avaliam seu impacto em domínios especializados.

Na área da saúde, RAG tem sido empregado para mitigar lacunas factuais em aplicações clínicas [15, 21]. Entretanto, aplicações voltadas a perguntas sobre medicamentos em português ainda são escassas, especialmente aquelas que analisam conjuntamente qualidade das respostas, custo computacional e avaliações por especialistas humanos e *LLM-as-a-Judge* [22].

Diferentemente dos trabalhos anteriores, investigamos essa questão em um domínio farmacêutico, comparando RAG convencional e QA-RAG quanto à recuperação semântica, qualidade das respostas, custo computacional e avaliação entre especialistas humanos e *LLM-as-a-Judge*.

3 METODOLOGIA EXPERIMENTAL

A Figura 1 resume o pipeline experimental. Inicialmente, coletamos as bulas dos 30 medicamentos mais vendidos no Brasil a partir da Farmarcas² e da Anvisa (Figura 1(a)). As bulas foram convertidas para texto, segmentadas em *chunks* e utilizadas para gerar 2647 perguntas sintéticas (Figura 1(b–c)). Em seguida, uma *LLM-as-a-Judge* (GPT) gerou 90 perguntas de avaliação (Figura 1(d)), onde foi gerado as resposta pela LLM (Qwen) em todos três cenários: *Zero-shot*, RAG

convencional e QA-RAG (Figura 1(e)). Por fim, as respostas foram comparadas por especialistas humanos e por uma *LLM-as-a-Judge* (GPT) (Figura 1(f)). Utilizamos a Qwen3-4B para geração das perguntas sintéticas e das respostas, e o GPT-5.5-2026-04-23 para gerar as perguntas de avaliação e realizar a comparação entre as respostas.

3.1 Bulas dos medicamentos (Figura 1(a))

Selecionamos os 30 medicamentos mais vendidos no Brasil nos quatro primeiros meses de 2024 segundo a Farmarcas². As respectivas bulas destinadas a profissionais de saúde foram obtidas no bulário eletrônico da Anvisa³ e utilizadas como base documental dos experimentos.

3.2 Conversão e indexação (Figura 1(b–c))

As bulas foram convertidas para texto com o Docling [12], segmentadas em *chunks* e indexadas de duas formas. No RAG convencional, recuperamos os cinco *chunks* semanticamente mais similares à consulta. No QA-RAG, geramos automaticamente uma pergunta sintética para cada *chunk* utilizando o prompt do Quadro 1; durante a inferência, a recuperação foi realizada sobre essas perguntas. Em todos os experimentos, utilizamos temperatura 0 (que busca deixar a geração de texto determinístico) e respostas limitadas a duas frases.

Quadro 1: Prompt utilizado para geração de perguntas da base documental, utilizada no QA-RAG.

system
 Você é um sistema de extração de informação especializado em farmacologia.
 Sua tarefa é gerar pares de PERGUNTA e RESPOSTA com base EXCLUSIVA no texto fornecido.
 REGRAS OBRIGATÓRIAS:
 - Use apenas informações presentes no texto
 - Perguntas devem ser específicas e objetivas
 - Respostas devem ser curtas (máximo 2 frases)
 - Responda APENAS em JSON válido
 Formato obrigatório:
 "pergunta": "...", "resposta": "..."

user
 Gere 1 par de pergunta e resposta com base no texto abaixo, inclua o nome do medicamento [NOME MEDICAMENTO] nas perguntas.
 TEXTO: [CHUNK MEDICAMENTO]
 JSON:

Foram geradas 2647 perguntas convertidas em embeddings com o modelo Paraphrase-multilingual-MiniLM-L12-v2 (Sentence Transformers) e indexadas com Faiss [6], viabilizando a recuperação por similaridade semântica das perguntas e *chunks* mais relevantes.

3.3 Geração das perguntas de avaliação e respostas (Figura 1(d–e))

Para construir o conjunto de avaliação, utilizamos o GPT-5.5-2026-04-23 para gerar automaticamente três perguntas por medicamento, totalizando 90 perguntas relacionadas aos 30 medicamentos selecionados. O prompt empregado é apresentado no Quadro 2. Com o conjunto de perguntas definido, submetemos as 90 questões a uma mesma LLM em três cenários experimentais:

²<https://www.farmacas.com.br/>

³<https://consultas.anvisa.gov.br/#/bulario/>

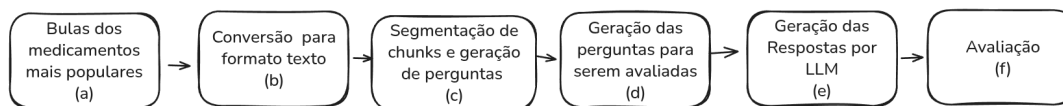


Figura 1: Fluxograma da proposta.

Quadro 2: Prompt utilizado para geração das perguntas de avaliação.

Você é um assistente de farmacologia que auxilia profissionais de saúde.
 Gere três perguntas para os medicamentos a seguir, assumindo que são perguntas que profissionais de saúde fariam ao chat:
 [NOME MEDICAMENTO]

(i) sem contexto adicional, caracterizando a abordagem *zero-shot*, e (ii) com contexto recuperado pelo mecanismo de RAG e (iii) com contexto recuperado pelo mecanismo de QA-RAG. O prompt utilizado para geração das respostas é apresentado no Quadro 3. Essa configuração permite isolar o impacto da recuperação semântica sobre a qualidade das respostas geradas.

Quadro 3: Prompt utilizado para geração das respostas.

system
 Você é um assistente de farmacologia.
 Responda em até 2 frases.
 user
 ## Pergunta:
 PERGUNTA PARA AVALIAÇÃO
 ## Contexto
 CONTEXTO RETORNADO PELO RAG ou QA-RAG
 ## Resposta:

3.4 Avaliação (Figura 1(f))

Por fim, as respostas foram comparadas por especialistas humanos e por uma abordagem baseada em *LLM-as-a-Judge*, cujo prompt é apresentado no Quadro 4. Embora apresente alta correlação com avaliações humanas [22], essa abordagem pode favorecer respostas mais detalhadas.

Quadro 4: Prompt utilizado para avaliação comparativa das respostas. Alternativa A utilizando Zero-Shot e B utilizando QA-RAG.

Você é um assistente de farmacologia.
 Avalie a melhor alternativa para cada pergunta abaixo:
 ## Pergunta X: Qual é a indicação principal do Naridrin e como ele atua no organismo?
 ## Alternativa A: Naridrin é indicado como antisséptico tópico para infecções bacterianas superficiais.
 ## Alternativa B: Naridrin é indicado para aliviar sintomas de rinite e congestão nasal, atuando como descongestionante nasal.

4 RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados organizados segundo as três questões de pesquisa. Inicialmente, avaliamos o impacto da recuperação de contexto na qualidade das respostas (RQ1). Em seguida, analisamos o custo computacional entre Zero-Shot e o QA-RAG (RQ2). Por fim, investigamos o efeito dessa estratégia sobre a recuperação semântica e a qualidade das respostas (RQ3).

Como referência, consideramos três cenários: (i) *Zero-shot*, sem recuperação de contexto; (ii) RAG convencional baseado em BM25 [14]; e (iii) RAG convencional baseado em similaridade de *embeddings*, utilizado como principal baseline para comparação com o QA-RAG.

RQ1: Qual é o impacto da recuperação de contexto na qualidade das respostas geradas em comparação ao Zero-shot?

Comparamos a abordagem *Zero-shot* com duas estratégias de recuperação: RAG convencional (BM25) e a abordagem proposta (QA-RAG). O objetivo foi verificar não apenas se a recuperação melhora as respostas, mas se diferentes estratégias trazem ganhos distintos.

Os resultados mostram que simplesmente adicionar recuperação não é suficiente. O RAG com BM25 empatou com o *Zero-shot* (45 vitórias para cada), indicando que similaridade lexical por termos não trouxe contexto suficientemente útil. Já o QA-RAG foi superior em 63 das 90 questões, contra 27 do *Zero-shot* — um aumento de 133% no número de vitórias, com diferença estatisticamente significativa ($p \approx 0.00019$).

Essa discrepância evidencia que o benefício da recuperação depende mais da **qualidade** do contexto do que da sua mera presença. Ao recuperar perguntas sintéticas semanticamente relacionadas, o QA-RAG seleciona trechos das bulas mais alinhados à intenção da consulta do usuário, reduzindo ambiguidades e alucinações.

O Quadro 4 ilustra bem essa diferença: enquanto o QA-RAG responde corretamente sobre o Naridrin, o *Zero-shot* alucina, descrevendo-o incorretamente como antisséptico antibacteriano.

Essa conclusão reforça que, em domínios especializados e documentos técnicos como bulas de medicamentos, recuperar por similaridade lexical pode ser insuficiente. Representar explicitamente as intenções dos usuários por meio de perguntas sintéticas revela-se uma estratégia mais eficaz.

RQ2: Qual é o custo computacional associado à substituição da indexação baseada em chunks por perguntas sintéticas?

Para responder a essa questão, analisamos separadamente os custos de indexação e de inferência do QA-RAG. Essa distinção é importante porque a geração das perguntas sintéticas ocorre apenas uma vez durante a construção da base documental, enquanto a etapa de recuperação é executada continuamente durante a utilização do sistema.

Os experimentos foram realizados em uma máquina equipada com processador Intel Xeon 2.20GHz, 32GB de memória RAM e GPU NVIDIA P100. A Tabela 1 apresenta o tempo médio gasto nas principais etapas do pipeline. O maior custo computacional concentra-se na geração das 2647 perguntas sintéticas (10080s), realizada apenas durante a fase de indexação. Em contraste, tanto a geração dos *embeddings* quanto a recuperação semântica apresentam custo praticamente desprezível (aproximadamente 1s).

Durante a inferência, a abordagem *Zero-shot* demandou aproximadamente 397s para responder às 90 perguntas, enquanto o QA-RAG consumiu cerca de 600s. Embora isso

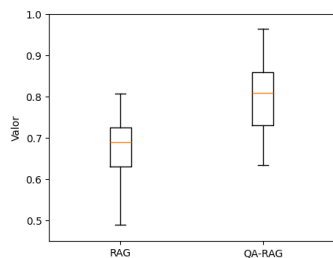
Tabela 1: Tempo médio de execução das diferentes etapas do pipeline seguido do intervalo de confiança de 95%.

Etapas	Tempo (s)
Geração de perguntas a partir dos <i>chunks</i>	10080 (149)
Geração de <i>embeddings</i> da base	1 (0.1)
Recuperação semântica (QA-RAG)	1 (0.1)
Geração de respostas com QA-RAG	600 (12)
Geração de respostas sem RAG (<i>Zero-shot</i>)	397 (3)

represente um aumento de aproximadamente 51% no tempo de resposta, esse acréscimo é relativamente pequeno quando comparado aos ganhos de qualidade observados na RQ1, na qual o QA-RAG obteve um aumento de 133% no número de respostas preferidas em relação ao cenário *Zero-shot*. Assim, o custo adicional de recuperação é amplamente compensado pela melhoria na qualidade das respostas produzidas.

Esses resultados mostram que o principal custo introduzido pelo QA-RAG está concentrado em uma etapa de pré-processamento executada apenas uma única vez. Após a construção da base de perguntas sintéticas, o índice pode ser reutilizado indefinidamente, fazendo com que esse custo seja amortizado ao longo de múltiplas consultas. Dessa forma, a substituição da indexação baseada em *chunks* por perguntas sintéticas caracteriza um *trade-off* favorável entre eficácia da recuperação e custo computacional, também explorado em estratégias de seleção eficiente de instâncias [2, 3], especialmente em aplicações nas quais a base documental permanece relativamente estável e é consultada repetidamente.

RQ3: Em que medida substituir a indexação baseada em *chunks* por perguntas sintéticas melhora a recuperação semântica e a qualidade das respostas? Inicialmente, avaliamos o impacto da unidade de indexação na etapa de recuperação. Enquanto o RAG convencional baseado em *embeddings* compara a consulta diretamente aos *chunks* das bulas, o QA-RAG recupera perguntas sintéticas previamente geradas a partir desses mesmos documentos. Como mostra a Figura 2, o QA-RAG apresenta escores de similaridade semanticamente superiores de forma consistente, indicando melhor alinhamento entre as consultas e os elementos recuperados.

**Figura 2: Distribuição dos escores de similaridade semântica obtidos pelo RAG convencional baseado em *embeddings* e pelo QA-RAG.**

Em seguida, analisamos se esse ganho na recuperação se reflete na qualidade das respostas. Na avaliação baseada em *LLM-as-a-Judge*, o QA-RAG foi considerado superior em 49 das 90 questões, contra 41 do RAG convencional. A mesma

tendência foi observada na avaliação realizada por dois especialistas da área da saúde, que preferiram o QA-RAG em 49 questões, o RAG convencional em 34 e registraram empate em sete casos. Há uma concordância entre humanos e o LLM-Judge em 65 casos. A Tabela 2 apresenta exemplos representativos dessas diferenças.

Os resultados indicam que a melhoria na qualidade das respostas decorre diretamente da recuperação de evidências mais alinhadas à intenção da consulta. Ao indexar perguntas sintéticas em vez de segmentos documentais, o QA-RAG representa explicitamente possíveis necessidades informacionais dos usuários, favorecendo uma recuperação semanticamente mais precisa e, consequentemente, respostas mais completas e consistentes. Esses resultados sugerem que a escolha da unidade de indexação constitui um fator relevante para sistemas RAG em domínios especializados.

Tabela 2: Exemplo qualitativo de comparação entre RAG convencional por *embedding* e QA-RAG. Em todos os exemplos, o especialista humano escolheu QA-RAG como a melhor resposta em relação a RAG.

Pergunta: Quais são os principais efeitos adversos do Glifage e qual é a complicação mais grave associada ao seu uso?

RAG: O Glifage é um medicamento utilizado no tratamento da diabetes tipo 2, e seus efeitos adversos principais incluem fadiga, tontura e vertigem. A complicação mais grave associada ao seu uso é a hipersensibilidade, que pode levar a reações anafiláticas e edema de face, lábios, faringe e/ou língua.

QA-RAG: Os principais efeitos adversos do Glifage incluem náusea, vômito, diarreia e distúrbios do paladar. A complicação mais grave associada ao seu uso é a acidose láctica, que pode ser fatal.

Pergunta: Quais são os riscos associados ao uso prolongado de Neosoro?

RAG: O uso prolongado de Neosoro pode causar irritação local e, em casos de superdosagem, pode levar a sintomas como sonolência e alodinia. Deve ser utilizado por um período não superior a 5 dias, conforme indicado..

QA-RAG: O uso prolongado de Neosoro pode causar náusea, cefaleia, congestão nasal por efeito rebote, rinite medicamentosa e hipertensão. O uso crônico pode também promover retenção urinária e sonolência..

5 CONCLUSÃO E TRABALHOS FUTUROS

Este estudo avaliou o uso de QA-RAG (indexação por perguntas sintéticas) em oposição aos *chunks* tradicionais no domínio farmacêutico (Anvisa). Constatou-se que o QA-RAG proporciona melhorias consistentes na recuperação e na qualidade das respostas, superando o RAG convencional nas avaliações humanas e automáticas. Embora haja maior custo na criação do índice, este é diluído no uso, gerando um *trade-off* positivo. Reitera-se que a ferramenta usa IA e não substitui avaliação clínica.

A escolha da unidade de indexação é crucial em sistemas RAG especializados. Como trabalhos futuros, buscaremos investigar novas estratégias de indexação e reranking, validar o sistema com perguntas reais, discutir os riscos de alucinações no contexto médico e expandir a base de medicamentos e modelos. Planejamos também analisar *clusters* [16], gerar perguntas sintéticas via reconhecimento de entidades [5], propor uma taxonomia de erros [1] e investigar a separabilidade dos *embeddings* [4].

CONSIDERAÇÕES ÉTICAS

Este estudo utilizou exclusivamente dados públicos (bulas da Anvisa), sem coletar informações sensíveis de pacientes. Devido aos riscos de alucinações de LLMs na área médica, ressalta-se que a ferramenta tem caráter estritamente experimental. As respostas geradas pela IA não constituem recomendações clínicas e não substituem a avaliação de um profissional de saúde.

AGRADECIMENTOS

Os autores declaram o uso de IA exclusivamente para revisão textual e execução dos modelos, assumindo total responsabilidade pelo conteúdo final. A solução técnica proposta foi desenvolvida empregando os modelos Qwen3-4B e GPT-5.5-2026-04-23. Este trabalho foi apoiado por: CNPq, CAPES, INCT-TILDIAR (Proc. 408490/2024-1), CIIA-Saúde (Proc. PPE-00030-21), Ministério da Saúde/SAPS (TED 31/2022, Proc. 25000.058753/2022-59), UNIMED-BH, FAPEMIG, AWS, NVIDIA e FAPESP.

REFERÊNCIAS

- [1] Claudio Moisés Valiense De Andrade, Washington Cunha, Guilherme Fonseca, Ana Clara Souza Pagano, Luana De Castro Santos, Adriana Silvina Pagano, Leonardo Chaves Dutra Da Rocha, and Marcos André Gonçalves. 2024. Explaining the Hardest Errors of Contextual Embedding Based Classifiers. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, Libby Barak and Malihe Alikhani (Eds.). Association for Computational Linguistics, Miami, FL, USA, 419–434. doi:10.18653/v1/2024.conll-1.31
- [2] Washington Cunha, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2023. An Effective, Efficient, and Scalable Confidence-Based Instance Selection Framework for Transformer-Based Text Classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 665–674. doi:10.1145/3539618.3591638
- [3] Washington Cunha, Felipe Viegas, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2025. A Noise-Oriented and Redundancy-Aware Instance Selection Framework. *ACM Transactions on Information Systems* 43, 1 (2025). doi:10.1145/3701769
- [4] Claudio M.V. de Andrade, Fabiano M. Belém, Washington Cunha, Celso França, Felipe Viegas, Leonardo Rocha, and Marcos André Gonçalves. 2023. On the class separability of contextual embeddings representations – or “The classifier does not matter when the (text) representation is so good!”. *Information Processing Management* 60, 4 (2023), 103336. doi:10.1016/j.ipm.2023.103336
- [5] Claudio M. V. de Andrade, Fabiano Muniz Belém, Celso França, Marcos Carvalho, Marcelo Ganem, Gabriel Teixeira, Gabriel Jalais, Alberto H. F. Laender, and Marcos A. Gonçalves. 2025. PromptNER: An Automatic Prompt-Learning Data Labeling Approach for Named Entity Recognition on Sensitive Data. *Journal of Information and Data Management* 16, 1 (Jan. 2025), 62–71. doi:10.5753/jidm.2025.4298
- [6] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]
- [7] David Fernandes, Edleno Silva de Moura, Berthier A. Ribeiro-Neto, Altigran Soares da Silva, and Marcos André Gonçalves. 2007. Computing Block Importance for Searching on Web Sites. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*. ACM, 165–174. doi:10.1145/1321440.1321466
- [8] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1762–1777. doi:10.18653/v1/2023.acl-long.99
- [9] Yunfan Gao et al. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2023). <https://arxiv.org/abs/2312.10997>
- [10] Seokgi Lee. 2025. Transforming Questions and Documents for Semantically Aligned Retrieval-Augmented Generation. arXiv:2508.09755 [cs.CL]
- [11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates.
- [12] Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2025. Dolving: An Efficient Open-Source Toolkit for AI-driven Document Conversion. arXiv:2501.17887 [cs.CL]
- [13] Vatsal Raina and Mark Gales. 2024. Question-Based Retrieval using Atomic Units for Enterprise RAG. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 219–233. <https://aclanthology.org/2024.feve-1.25/>
- [14] Stephen Edward Robertson, Steve Walker, Susan Jones, Michelle M Hancock-Beaulieu, Mike Gatford, et al. 1994. Okapi at TREC. (1994).
- [15] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Ting, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.
- [16] Felipe Viegas, Sergio Canuto, Washington Cunha, Celso França, Claudio Valiense, Leonardo Rocha, and Marcos Gonçalves. 2023. CluSent – Combining Semantic Expansion and De-Noising for Dataset-Oriented Sentiment Analysis of Short Texts. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web (Ribeirão Preto/SP)*. SBC, Porto Alegre, RS, Brasil.
- [17] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Cunha, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 753–761. doi:10.1145/3289600.3291032
- [18] Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos André Gonçalves. 2020. CluHTM: Semantic Hierarchical Topic Modeling Based on CluWords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 8138–8150.
- [19] Felipe Viegas, Washington Cunha, Christian Gomes, Leonardo Rocha, and Marcos André Gonçalves. 2025. Exploiting Contextual Embeddings in Hierarchical Topic Modeling and Investigating the Limits of the Current Evaluation Metrics. *Computational Linguistics* (2025).
- [20] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9414–9423. doi:10.18653/v1/2023.emnlp-main.585
- [21] Cyril Zakra, Akash Chaurasia, Rohan Shad, Alex R Dalal, Jennifer L Kim, Michael Moor, Kevin Eng, Christopher Fong, William Phillips, Michelle M Mello, et al. 2024. Almanac: retrieval-augmented language models for clinical medicine. *NEJM AI* 1, 2 (2024), A10a2300068.
- [22] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Hao, Zhanghao Wu, Sen Ba, Zi Lin Jiang, Hao Wu, Yonghao Zhuang, Zi Lin, et al. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, Vol. 36.