Retrieval-Augmented Generation for Clinical Question Answering in Portuguese Drug Leaflets: Benefits and Limitations

Gabriel Lino Garcia[1], Pedro Henrique Paiola[1], João Vitor Mariano Correia[1], Douglas Rodrigues[1], João Paulo Papa[1],

[1] São Paulo State University (UNESP)
Av. Eng. Luís Edmundo Carrijo Coube, 14-01 - Bauru - SP - Brazil ,
Correspondence: gabriel.lino@unesp.br

Abstract

Retrieval-Augmented Generation (RAG) is proposed to reduce hallucination and improve grounding in clinical language models, yet its effectiveness across different levels of clinical reasoning remains unclear. We conducted a controlled evaluation of medication-related question answering in Portuguese using over 7,000 Brazilian regulatory drug leaflets and a complementary clinical benchmark derived from national medical licensing examinations (Revalida and Fuvest). Retrieval substantially improved factual recall and clinical coherence in medication-specific queries, increasing F1 from 0.276 to 0.412. However, naive retrieval did not consistently improve complex clinical reasoning and sometimes reduced accuracy compared to a parametric-only baseline. We identify retrieval-induced anchoring bias, where partially relevant evidence shifts model decisions toward clinically incorrect conclusions. Critique-based and adaptive retrieval mitigated this effect and achieved the highest clinical benchmark accuracy (54.25%). Clinically grounded evaluation dimensions revealed safety-relevant differences beyond traditional NLP metrics. These results show that retrieval augmentation is effective in regulatory settings but requires adaptive control for higher-level clinical reasoning.

1 Introduction

Access to reliable and up-to-date medication information remains a persistent challenge in clinical practice, particularly in low- and middle-income countries such as Brazil. Drug leaflets issued by regulatory agencies constitute the official, legally binding source of information on indications, contraindications, dosage, pharmacodynamics, and adverse effects. However, their length, technical density, and fragmented organization frequently hinder efficient information retrieval by healthcare professionals and patients. Difficulties navigating regulatory documentation may increase reliance on secondary summaries or informal sources, which are not always complete, up to date, or clinically safe.

In recent years, large language models (LLMs) have demonstrated remarkable performance in medical question answering, clinical reasoning, and decision support tasks (Singhal et al., 2023). Despite these advances, LLMs operate primarily through parametric knowledge acquired during large-scale pretraining (Brown et al., 2020). When deployed without explicit grounding mechanisms, they are prone to hallucinations, outdated recommendations, and unverifiable justifications (Ji et al., 2023). These limitations are particularly concerning in high-stakes medical environments, where incorrect or poorly supported outputs may lead to harmful clinical decisions. Empirical studies have shown that fluent responses generated by LLMs can contain subtle but clinically significant inaccuracies, reinforcing concerns regarding their safe adoption in healthcare settings (Singhal et al., 2023).

Large language models have recently approached expert-level performance on standardized clinical benchmarks, as illustrated by Med-PaLM 2, which demonstrated substantial improvements in accuracy and clinician-preferred responses on multiple medical QA datasets (Singhal et al., 2025). Such advances highlight LLMs' growing ability to handle complex medical questions, but they also underscore the need for grounded, reliable outputs that align with clinical evidence.

Retrieval-Augmented Generation (RAG) has emerged as a principled strategy to mitigate these risks by conditioning generation on externally retrieved documents (Lewis et al., 2020). By incorporating relevant evidence at inference time, RAG systems aim to improve factual accuracy, traceability, and transparency. In biomedical and clinical domains, retrieval-based augmentation has been applied to clinical guidelines, scientific literature, and electronic health records, often yielding im-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 112–120

provements in evidence attribution and factual consistency (Xiong et al., 2024). Nevertheless, the prevailing narrative that retrieval universally enhances performance has not been rigorously examined in settings that require higher-order reasoning or contextual abstraction.

This limitation becomes particularly salient in medication-related question answering. Drug leaflets are authoritative regulatory documents that provide structured and legally validated information. However, they are not designed to support complex therapeutic reasoning, differential diagnosis, or integrated clinical decision-making. When retrieval mechanisms introduce partial, excessive, or poorly aligned context, the model may anchor its reasoning to suboptimal evidence. We refer to this phenomenon as *retrieval-induced anchoring bias*. In such cases, responses may remain formally grounded in retrieved passages while still being clinically inappropriate or incomplete. Despite the importance of this issue for patient safety, systematic empirical analyses of retrieval-induced bias in medical QA remain scarce.

The challenge is amplified in Portuguese, which is among the most widely spoken languages globally and the primary language of clinical practice in Brazil. Most studies on medical question answering and retrieval-augmented systems focus predominantly on English-language resources (Singhal et al., 2023). Despite the growing adoption of NLP in healthcare, Portuguese clinical resources remain comparatively scarce. Prior work has focused primarily on foundational components such as domain-specific embeddings (e Oliveira et al., 2019), clinical information extraction from medical records (Da Rocha et al., 2022), and semantically annotated corpora such as SemClinBr (Oliveira et al., 2022).

Specialized biomedical language models have also been developed for Portuguese. For Brazilian Portuguese, BioBERTpt was introduced to support clinical named entity recognition tasks (Schneider et al., 2020), while MediAlbertina is a recent large-scale medical language model for European Portuguese (Nunes et al., 2024). More recent efforts have begun adapting large language models to the medical domain in Portuguese, including studies on fine-tuning strategies and domain transfer (Paiola et al., 2024). In parallel, dedicated benchmarks for Brazilian Portuguese medical NLP have recently emerged, such as the *DrBodeBench* benchmark proposed by (Garcia et al., 2025), highlighting the lack of standardized evaluation resources for clinical reasoning tasks.

However, these resources largely target entity-level understanding or document processing tasks rather than complex clinical reasoning or evidence-grounded question answering. Consequently, publicly available Portuguese benchmarks for clinical QA remain limited, and evaluation frameworks rarely incorporate clinically meaningful dimensions beyond lexical overlap metrics. Standard NLP measures such as Exact Match and F1 do not fully capture groundedness, hallucination risk, or potential clinical harm, dimensions that are increasingly emphasized in trustworthy artificial intelligence research (Rudin, 2019; Ji et al., 2023).

In this work, we conduct a systematic, controlled evaluation of Retrieval-Augmented Generation for medication-related clinical question answering in Portuguese, using Brazilian regulatory drug leaflets as the primary source of authoritative knowledge. Our study is designed to disentangle the role of retrieval across different levels of clinical abstraction. To this end, we evaluate RAG across two complementary, carefully constructed settings. The first setting comprises a controlled, medication-specific QA dataset derived directly from curated regulatory leaflets, with questions tightly aligned with the explicit document content. The second setting consists of a broader clinical reasoning benchmark built from a subset of the Portuguese medical benchmark DrBodeBench (Garcia et al., 2025), which includes questions from Brazilian medical examinations such as Revalida and the FUVEST direct-access residency exam, where medication knowledge must be integrated into complex diagnostic, therapeutic, and decision-making scenarios.

The main contributions of this paper are summarized as follows:

- We introduce a curated Portuguese question answering dataset derived from Brazilian drug leaflets. The dataset includes extensive preprocessing to address crawler-induced extraction noise, structural inconsistencies, and metadata imprecision.

- We construct a complementary clinical QA benchmark by systematically identifying and extracting medication-related questions from the Brazilian medical licensing examination, thereby enabling evaluation in realistic and high-stakes clinical reasoning settings.

- We provide a comprehensive empirical comparison of six modeling strategies, ranging from a base language model without retrieval to multiple retrieval-augmented variants, including fusion-based, hypothetical-document, critique-based, and adaptive approaches.
- We develop a clinically oriented evaluation framework that integrates traditional NLP metrics with groundedness verification, hallucination detection, and explicit clinical risk assessment through an LLM-as-a-judge protocol.
- We offer empirical evidence that Retrieval-Augmented Generation is not universally beneficial in medical question answering and demonstrate that adaptive retrieval control is critical for safe and reliable deployment in complex clinical scenarios.

Overall, our findings contribute to a more nuanced and safety-aware understanding of Retrieval-Augmented Generation in healthcare. We characterize the conditions under which retrieval enhances factual reliability, identify scenarios in which it may introduce reasoning bias, and provide actionable guidance for the responsible development of grounded clinical language models in Portuguese and other underrepresented linguistic contexts.

The remainder of this paper is organized as follows. Section 2 describes the construction and characteristics of the curated drug leaflet dataset and the DrBodeBench-derived medication-focused clinical benchmark. Section 3 details the modeling strategies, including the base language model and its retrieval-augmented variants. Section 4 presents the experimental setup, evaluation protocol, quantitative results, and a comprehensive discussion of clinical safety implications, observed failure modes, and methodological limitations. Finally, Section 5 concludes the paper and outlines directions for future research.

2 Datasets and Benchmark Design

To systematically analyze the effectiveness of Retrieval-Augmented Generation under varying levels of clinical abstraction, we constructed two complementary evaluation benchmarks. The first benchmark isolates document-grounded factual retrieval in a controlled regulatory setting. The second benchmark evaluates retrieval behavior in realistic, high-level clinical reasoning scenarios derived from national medical licensing examinations. This dual design enables a structured investigation of how question specificity and abstraction level modulate retrieval effectiveness.

2.1 Bulário Regulatory Corpus

Our primary knowledge source is the publicly available *Bulário* corpus (Cunha et al., 2018), which contains over 7,000 Brazilian regulatory drug leaflets. These documents constitute the official legal source of information regarding indications, contraindications, dosage, pharmacodynamics, adverse reactions, and drug interactions.

The raw corpus contained significant structural noise introduced by automated crawling during its original compilation. We therefore implemented a multi-stage preprocessing pipeline:

- Removal of duplicated, truncated, or incomplete leaflets;
- Correction of HTML parsing artifacts and formatting inconsistencies;
- Standardization of section headers to enable consistent document segmentation;
- Normalization of metadata fields, including medication names, active substances, and therapeutic classes.

After cleaning, the corpus was indexed using BM25 for lexical retrieval. Preliminary experiments indicated that purely embedding-based semantic retrieval was unstable due to domain-specific terminology and heterogeneous document structure. BM25 provided more reliable alignment with regulatory language and section-level content.

2.2 Medication-Specific QA Benchmark

To construct a controlled evaluation benchmark, we selected 25 widely prescribed medications in Brazil across four therapeutic categories: antibiotics, analgesics and anti-inflammatory agents, antihypertensives, and antidiabetics. These categories were chosen to ensure clinical diversity across infectious, inflammatory, cardiovascular, and metabolic conditions.

For each selected medication, we verified the presence of its corresponding leaflet in the cleaned corpus. We then identified 10 standardized sections consistently present across documents, including indications, dosage, contraindications, warnings, adverse reactions, and drug interactions.

Question generation was performed using a large language model conditioned on the medication name and structured leaflet content. To ensure balanced coverage and section-level control, we generated:

- 4 questions per medication per section;
- 40 questions per section across medications;
- 400 total open-ended questions.

This benchmark, referred to as the **Medication-Specific QA Benchmark**, is explicitly designed to measure factual recall, section-level grounding, and evidence attribution when answers are directly localized within regulatory documents[1].

2.3 Medication-Focused Clinical Benchmark from DrBodeBench

To evaluate retrieval capabilities in higher-level reasoning scenarios, we created a second benchmark derived from the Portuguese medical benchmark DrBodeBench (Garcia et al., 2025). This benchmark aggregates questions from Brazilian medical examinations, including the Revalida and the FUVEST direct-access residency exam. From DrBodeBench, we curated a specific subset of questions that exclusively pertains to medication-related topics.

A large language model was employed to identify examination items in which medication knowledge plays a central role in diagnostic or therapeutic decision-making. Only questions requiring explicit pharmacological integration were retained. The resulting dataset consists of clinically contextualized multiple-choice scenarios that require integration of medication knowledge with patient history, laboratory findings, and clinical reasoning.

In contrast to the controlled leaflet-based benchmark, answers in this dataset are not necessarily localized within a single document section. Instead, they frequently require synthesis across distributed knowledge and contextual interpretation. This property makes the benchmark suitable for analyzing potential retrieval-induced bias in complex reasoning tasks[2].

[1] https://huggingface.co/datasets/recogna-nlp/bulas_qa

[2] https://huggingface.co/datasets/recogna-nlp/drbodebench_medicamentos

3 Methodology

This section describes the modeling framework used to evaluate RAG for medication-related clinical question answering in Portuguese. The study is designed around a single core hypothesis: the effect of retrieval depends on question specificity and on the level of clinical abstraction required. To test this hypothesis, we compare a parametric-only language model against multiple retrieval-augmented variants on two complementary benchmarks, one dominated by localized, document-explicit answers and another requiring higher-level clinical reasoning.

To enable controlled comparisons, we keep the underlying generator fixed across all methods and vary only the retrieval and evidence-integration procedures. Such a controlled setup isolates performance differences attributable to retrieval behavior rather than to model capacity.

3.1 Task Formulation

Let q denote a clinical question and let $\mathcal{D}$ denote the cleaned Bulário collection indexed for retrieval. A retriever R maps q to a ranked list of passages

$$P_K = \{p_1, \ldots, p_K\}, \quad P_K \subset \mathcal{D},$$

where each p_i is a chunk extracted from a drug leaflet, a generator G produces an answer a either from parametric knowledge alone or conditioned on retrieved evidence:

$$a = \begin{cases} G(q) & \text{(parametric-only)}, \\ G(q, P_K) & \text{(retrieval-augmented)}. \end{cases}$$

This formulation supports direct comparisons between parametric-only reasoning and evidence-grounded generation under identical decoding settings. It also allows evaluation across two regimes: medication-specific questions, for which the answer is usually explicitly stated in the leaflet, and complex clinical reasoning questions from the DrBodeBench subset, for which the answer often requires integrating pharmacological knowledge with broader clinical context.

3.2 Evidence Preparation and Retrieval Backbone

Drug leaflets were segmented into section-aware chunks based on standardized headers. Each chunk retained metadata that preserves its provenance, including medication name and section identifier.

This design supports section-level retrieval analysis and enables evaluation of retrieval coverage by leaflet section.

We employ BM25 as the primary retrieval method. In preliminary experiments, purely embedding-based semantic retrieval produced unstable rankings in Portuguese regulatory text, likely due to specialized terminology and heterogeneous formatting. BM25 yielded more robust lexical alignment for this domain and served as the retrieval backbone for all RAG variants.

Unless stated otherwise, all retrieval-augmented methods use a fixed retrieval depth K to ensure comparability across systems.

3.3 Compared Systems

We evaluate six modeling strategies: a parametric-only baseline and five retrieval-augmented methods. All methods use the same generator G and differ only in how they retrieve, filter, or integrate evidence.

3.3.1 Parametric-only Baseline

The **Base** system produces an answer without external evidence:

$$a_{\text{base}} = G(q).$$

This baseline quantifies the performance achievable from parametric knowledge alone and provides a reference point for assessing both gains and degradations introduced by retrieval.

3.3.2 Direct Retrieval-Augmented Methods

RAG-Simple augments generation with the top-K retrieved chunks:

$$P_K = R_{\text{BM25}}(q), \quad a_{\text{simple}} = G(q, P_K).$$

This system tests whether direct evidence injection improves factual accuracy and grounding in medication-specific queries.

RAG-Fusion expands retrieval coverage by issuing multiple query reformulations and aggregating retrieved candidates before selecting the final evidence set P_K. This approach aims to reduce lexical mismatch and improve recall when relevant information is distributed across sections.

RAG-HyDE retrieves evidence using a hypothetical document produced by the generator. The system first generates a synthetic passage $\tilde{d}$ from the question, then retrieves evidence conditioned on $\tilde{d}$, and finally answers using the retrieved context:

$$\tilde{d} = G(q), \;\; P_K = R_{\text{BM25}}(\tilde{d}), \;\; a_{\text{hyde}} = G(q, P_K).$$

HyDE is intended to bridge the gap between the question phrasing and regulatory language, potentially improving retrieval precision.

These direct RAG variants are expected to be most effective when answers are explicitly stated in the documents. In higher-abstraction questions, however, they may introduce partial or poorly prioritized evidence that biases the selection of answers.

3.3.3 Critique-based and Adaptive Retrieval Methods

CRAG introduces an additional validation stage that assesses whether the retrieved context is adequate for the question. When evidence is judged insufficient or misaligned, CRAG triggers a refinement step that updates retrieval and re-generates the answer using improved context. This design targets failure modes in which naive retrieval anchors the model to incomplete evidence in clinically complex scenarios.

Adaptive RAG dynamically decides whether retrieval should be used and controls how much evidence is injected. Formally, a policy $\pi(q)$ selects whether to retrieve and the retrieval depth:

$$(\text{use\_retrieval}, K(q)) = \pi(q).$$

If retrieval is enabled, the system generates $a = G(q, P_{K(q)})$. Otherwise, it returns $a = G(q)$. This adaptive control is designed to mitigate retrieval-induced anchoring bias by reducing reliance on retrieved evidence when the question requires abstraction beyond what leaflets can directly support.

3.4 Design Rationale

The evaluated systems form a progression from parametric-only generation to increasingly controlled retrieval mechanisms. This design supports three empirical tests aligned with our claims: retrieval should improve grounding and factual recall for localized medication questions, naive retrieval may underperform in broader clinical reasoning tasks when evidence is partial or excessive, and adaptive or critique-based strategies should mitigate these failures by controlling when and how retrieval is used.

3.5 Evaluation Overview

All systems are evaluated under identical decoding settings across both benchmarks. We report standard lexical metrics such as Exact Match and F1, retrieval-sensitive metrics such as section-level Recall@K and grounding rate, and clinically oriented dimensions, including hallucination detection and clinical risk assessment, evaluated using the G-Eval framework (Liu et al., 2023), an LLM-based evaluation protocol designed to improve alignment with human judgments. We also report latency to quantify computational overhead introduced by multi-query retrieval and iterative evidence construction.

The full evaluation protocol and results are presented in Section 4.

4 Results and Discussion

The experimental results reveal a consistent, theoretically meaningful pattern: the impact of Retrieval-Augmented Generation is strongly conditioned by question specificity and the level of clinical abstraction required. Retrieval substantially improves performance when answers are explicitly localized in regulatory documents. Its effect becomes more complex in higher-level clinical reasoning scenarios, where uncontrolled evidence injection may interfere with decision-making.

4.1 Results on the Medication-Specific QA Benchmark

Table 1 reports performance on the Medication-Specific QA Benchmark.

Across all retrieval-augmented variants, improvements over the parametric-only baseline are substantial and consistent. The Base model achieves F1 = 0.276 and G-Eval = 0.517, reflecting limited factual precision and reduced clinical coherence even in a restricted domain. All RAG-based methods increase F1 by approximately 0.12 to 0.14 and improve G-Eval by more than 0.20.

These gains confirm that when answers are explicitly stated in regulatory leaflets, retrieval augmentation fulfills its primary objective. It enhances evidence grounding, increases section-level recall, and improves justification quality. Automatic grounding rates and Recall@K values approach ceiling levels across RAG variants, indicating that retrieval coverage is not the principal limiting factor. Performance differences among retrieval methods arise primarily from generation quality and contextual synthesis rather than from evidence availability.

Latency analysis reveals a cost–performance trade-off. RAG-Fusion achieves the highest F1 but nearly doubles inference time. RAG-HyDE attains the highest G-Eval score but incurs the largest computational overhead. CRAG and Adaptive RAG maintain competitive quality while preserving moderate latency. In this extractive setting, simple retrieval offers the most favorable balance between efficiency and performance.

Overall, retrieval is highly effective in document-grounded medication QA, where answers are explicitly localized, and the regulatory structure aligns with the reasoning task.

4.2 Results on the DrBodeBench Clinical Benchmark

A markedly different pattern emerges in the DrBodeBench medication subset, summarized in Table 2.

Direct retrieval strategies do not consistently improve performance in this setting. RAG-Simple and RAG-Fusion slightly underperform the Base model, with degradations of 0.5-1.8 percentage points. RAG-HyDE performs equivalently to the parametric baseline.

Questions in this subset require integrating pharmacological knowledge with patient history, laboratory findings, and therapeutic prioritization. Retrieved leaflet passages often contain accurate but partial information. When such fragments are injected without filtering or abstraction control, the generator may anchor on salient yet incomplete evidence and select suboptimal answer choices.

A comprehensive qualitative analysis presented in Table 3 reveals insights that accuracy alone does not capture. In a representative case involving varicella prophylaxis for an immunosuppressed child, the base model selected the correct answer based on parametric knowledge; however, it lacked explicit supporting evidence, leading the automated judge to classify it as ungrounded.

In contrast, RAG-based methods such as RAG-Simple and RAG-HyDE produced correct answers supported by excerpts from package inserts, which resulted in grounded justifications and reduced clinical risk. On the other hand, RAG-Fusion provided a clinically incorrect answer, despite being grounded in the retrieved content, leading to a classification of grounded and non-hallucinatory, but with high clinical risk.

This case underscores that grounding alone does

Table 1: Performance on the Medication-Specific QA Benchmark. Best results per metric are shown in bold.

Model	F1	G-Eval	Latency (s)
Base (no retrieval)	0.276	0.517	4.41
RAG-Simple	0.411	0.723	4.88
RAG-Fusion	**0.412**	0.723	8.86
RAG-HyDE	0.393	**0.735**	11.64
CRAG	0.393	0.728	5.50
Adaptive RAG	0.401	0.733	5.16

Table 2: Accuracy on the DrBodeBench Medication-Focused Clinical Benchmark. Best result is shown in bold.

Method	Accuracy (%)
Base (no retrieval)	51.85
RAG-Simple	51.36
RAG-Fusion	50.08
RAG-HyDE	51.85
CRAG	**54.25**
Adaptive RAG	52.81

not guarantee clinical accuracy; the retrieval of partially relevant or misleading information can lead the model to make erroneous therapeutic decisions. This behavior exemplifies retrieval-induced anchoring bias, in which retrieved evidence constrains reasoning, even when it does not support the correct answer.

These findings suggest that, for complex clinical reasoning tasks like those in the DrBodeBench, more selective or adaptive retrieval strategies are crucial to mitigate risks. Conversely, direct retrieval remains most effective for straightforward question-answering scenarios, such as those involving package inserts.

Overall, the example underscores the importance of evaluation frameworks that distinguish groundedness from clinical validity.

CRAG achieves the highest accuracy at 54.25%, and Adaptive RAG also surpasses the parametric baseline. Both methods regulate evidence exposure through critique or dynamic control, reducing reliance on misaligned context.

Taken together, the results demonstrate that Retrieval-Augmented Generation is highly effective in document-explicit medication QA but requires adaptive control in higher-abstraction clinical reasoning scenarios to ensure robustness and patient safety.

5 Conclusions and Future Directions

This work presents a systematic evaluation of Retrieval-Augmented Generation for medication-related clinical question answering in Portuguese, grounded in the cleaned Bulário regulatory corpus and validated on a complementary medical licensing benchmark derived from Revalida examinations. By structuring evaluation across two benchmarks that differ in question specificity and abstraction level, the study demonstrates that retrieval effectiveness is not universal. Instead, it depends critically on the specificity of the question and the level of clinical abstraction required.

In document-grounded medication-specific queries, retrieval substantially improves factual recall, evidence grounding, and clinical coherence. When answers are localized within regulatory leaflets, retrieval augmentation enhances reliability and traceability without meaningful performance trade-offs. These findings reinforce the value of RAG architectures in regulatory and extractive clinical information settings, particularly in Portuguese, where high-quality structured benchmarks remain limited.

In contrast, derived from Brazilian medical examination questions included in DrBodeBench, a more nuanced picture emerges. Direct retrieval strategies can slightly degrade performance by introducing partial or misaligned evidence that influences answer selection. This phenomenon, characterized in this study as retrieval-induced anchoring bias, highlights a critical limitation of naive context injection. Grounded responses are not necessarily clinically correct, and the presence of evidence alone does not guarantee safe reasoning.

More advanced retrieval strategies, including critique-based and adaptive approaches, consistently mitigate these limitations. By regulating when and how external evidence is incorporated, these methods improve robustness in clinically complex scenarios and reduce high-risk errors. The

Table 3: Illustrative qualitative example from the DrBodeBench subset.

Method	Grounded	Correct	Clinical Risk
Base	No	Yes	Low
RAG-Simple	Yes	Yes	None
RAG-HyDE	Yes	Yes	None
RAG-Fusion	Yes	No	High
CRAG	Yes	Yes	None

findings suggest that adaptive retrieval control is essential for safe deployment of language models in high-stakes medical environments.

Beyond performance metrics, the study underscores the importance of clinically oriented evaluation frameworks. Standard NLP metrics such as Exact Match and F1 fail to capture safety-relevant differences between systems. Dimensions such as groundedness, hallucination rate, and clinical risk provide complementary insights indispensable to healthcare applications.

Several avenues for future research emerge from these findings. Formal modeling of retrieval-induced anchoring bias could deepen understanding of how exposure to evidence shifts decision boundaries in generative models. Hybrid retrieval mechanisms combining lexical and semantic strategies may improve robustness in morphologically rich languages such as Portuguese. Policy-learning approaches could dynamically calibrate retrieval depth based on clinical risk. Finally, prospective validation in clinical workflows would strengthen the external validity of retrieval-augmented systems.

In summary, Retrieval-Augmented Generation is a powerful but context-dependent approach for clinical question answering. Its benefits are pronounced in document-explicit regulatory domains, yet careful control is required when reasoning extends beyond explicit evidence. For medical AI systems in Portuguese and other underrepresented languages, adaptive retrieval policies are not merely enhancements but foundational requirements for trustworthy and safe deployment.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Alexandre Cunha, Gabriel dos Santos, and Gustavo Paiva Guedes. 2018. Uma análise sobre as bulas de medicamentos no brasil. In *CSBC 2018 - 12º BreSci ()*.

Naila Camila Da Rocha, Abner Macola Pacheco Barbosa, Yaron Oliveira Schnr, Juliana Machado-Rugolo, Luis Gustavo Modelli de Andrade, José Eduardo Corrente, and Liciana Vaz de Arruda Silveira. 2022. Natural language processing to extract information from portuguese-language medical records. *Data*, 8(1):11.

Lucas Emanuel Silva e Oliveira, Yohan Bonescki Gumiel, Arnon Bruno Ventrilho Dos Santos, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid A Hasan, and Claudia Maria Cabral Moro. 2019. Learning portuguese clinical word embeddings: A multi-specialty and multi-institutional corpus of clinical narratives supporting a downstream biomedical task. In *MedInfo*, pages 123–127.

Gabriel Lino Garcia, João Renato Ribeiro Manesco, Pedro Henrique Paiola, Pedro Henrique Crespan Ribeiro, Ana Lara Alves Garcia, and João Paulo Papa. 2025. A step forward for medical llms in brazilian portuguese: Establishing a benchmark and a strong baseline. In *Proceedings of the 38th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2025)*, Madrid, Spain.

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).

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*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval:

NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Miguel Nunes, João Boné, João C Ferreira, Pedro Chaves, and Luis B Elvas. 2024. Medialbertina: An european portuguese medical language model. *Computers in Biology and Medicine*, 182:109233.

Lucas Emanuel Silva e Oliveira, Ana Carolina Peters, Adalniza Moura Pucca Da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022. Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

Pedro Henrique Paiola, Gabriel Lino Garcia, João Renato Ribeiro Manesco, Mateus Roder, Douglas Rodrigues, and João Paulo Papa. 2024. Adapting llms for the medical domain in portuguese: A study on fine-tuning and model evaluation. *Preprint*, arXiv:2410.00163.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd clinical natural language processing workshop*, pages 65–72.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.