Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BM25 and Dense Retrieval Are Complementary for Portuguese Clinical Text: An Empirical Study of Hybrid RAG Across 500 Clinical Queries

Igor Eduardo

igorcnnbd@gmail.com

independent researcher https://orcid.org/0009-0005-6288-1135

Case Report

Keywords: retrieval-augmented generation, clinical decision support, BM25, hybrid retrieval, Portuguese medical text, citation verification, LLM-as-judge

Posted Date: August 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10757954/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: The author is an AI Applied Engineer and founder of DocMinds AI, a clinical AI company whose production retrieval system was used as the evaluation environment for this study. The author therefore holds a financial interest in the system under evaluation.

BM25 and Dense Retrieval Are Complementary for Portuguese Clinical Text: An Empirical Study of Hybrid RAG Across 500 Clinical Queries

Igor Eduardo
ORCID: 0009-0005-6288-1135
igoreduardo.com github.com/nomad-link-id

Abstract

How should retrieval-augmented generation (RAG) systems be configured for clinical decision support in Portuguese? Despite growing deployment of LLM-based clinical tools in Latin America, no empirical study has evaluated hybrid retrieval strategies for Portuguese medical text. We present an evaluation of 500 clinical queries across 6 medical specialties against a production clinical corpus, comparing BM25, dense, and hybrid retrieval with Reciprocal Rank Fusion (RRF). Four findings emerge: (1) BM25 and hybrid retrieval surface statistically distinct document sets (McNemar χ^2=39.55, $p < 0.001$), with 680 sources unique to hybrid and 467 unique to BM25—confirming complementarity rather than redundancy; (2) dense-only retrieval fails for 22.2% of clinical queries even at the most permissive threshold, making it structurally unsuitable as the sole retrieval path; (3) authority-weighted scoring affects document ranking but not recall, suggesting that evidence hierarchy boosts should be evaluated by ranking quality rather than coverage metrics; (4) inter-annotator agreement for clinical relevance judgments between two independent LLM judges reaches κ=0.954, validating automated evaluation methodology for Portuguese medical text. Deterministic citation verification eliminates hallucinated references entirely (461/500 correct vs. 1/500 for prompt-based, Fisher's exact $p < 0.001$). These findings inform retrieval architecture decisions for clinical RAG systems operating in Portuguese and potentially other non-English clinical domains.

Keywords: retrieval-augmented generation, clinical decision support, BM25, hybrid retrieval, Portuguese medical text, citation verification, LLM-as-judge

1 Introduction

Retrieval-Augmented Generation has become the standard architecture for deploying large language models in knowledge-intensive clinical applications [1, 2]. However, the retrieval component is typically designed and evaluated for English medical text, and the assumption that dense semantic retrieval outperforms sparse keyword matching—well-established for English [3]—has not been empirically validated for Portuguese clinical terminology.

This gap matters. Brazil has 575,000+ physicians and a growing ecosystem of clinical AI tools operating in Portuguese. Retrieval failures in clinical systems carry patient safety implications: a missed guideline or a hallucinated citation can lead to incorrect clinical decisions [4]. Yet no published study evaluates how BM25, dense retrieval, and hybrid approaches compare for Portuguese clinical queries against a curated medical corpus.

We address this gap with four research questions:

RQ1: Do BM25 and hybrid retrieval surface the same or different document sets for Portuguese clinical queries?

RQ2: At what point does dense-only retrieval become unreliable for clinical use?

RQ3: Does authority-weighted scoring (boosting national protocols over primary studies) improve recall, ranking, or both?

RQ4: Can automated LLM-based relevance judgments substitute for physician annotation in evaluating Portuguese clinical retrieval?

2 Related Work

2.1 Sparse vs. Dense Retrieval in Biomedical Domains

Ma et al. [5] demonstrated that BM25 remains competitive with dense retrieval for entity-centric biomedical queries in English. Thakur et al. [6] showed that dense models trained on general corpora underperform BM25 on domain-specific benchmarks. Our work extends these findings to Portuguese clinical text, where terminology overlap between queries and protocol documents creates a distinct retrieval dynamic.

2.2 Reciprocal Rank Fusion

Cormack et al. [7] introduced RRF for merging heterogeneous retrieval signals. Chen et al. [8] showed RRF outperforms learned fusion with limited training data. We adopt RRF with k=60 and evaluate whether the fusion provides measurable benefit over BM25-alone for Portuguese clinical queries—a question not previously addressed.

2.3 Citation Verification in RAG

Gao et al. [9] reduced citation hallucinations to 3–5% using a secondary LLM verifier. Min et al. [10] introduced FActScore for fine-grained attribution evaluation. We compare prompt-based and deterministic citation verification, finding that exact document-array matching eliminates hallucinations entirely.

2.4 LLM-as-Judge for Retrieval Evaluation

Zheng et al. [11] established LLM-as-judge methodology for English. Thomas et al. [12] validated it for retrieval evaluation. No study has validated this methodology for Portuguese clinical text. We provide the first inter-annotator agreement measurement (κ) for LLM-based clinical relevance judgments in Portuguese.

2.5 Evidence Hierarchies in Clinical Retrieval

The GRADE framework [13] establishes evidence quality levels. Zakka et al. [2] incorporated guideline-level retrieval in Almanac. Our authority-weighting approach operationalizes GRADE-like hierarchies as retrieval boosts—and our evaluation reveals the counterintuitive finding that these boosts affect ranking without improving recall.

3 Method

3.1 Clinical Corpus

The evaluation corpus consists of curated clinical content spanning clinical guidelines, drug information, research literature, administrative coding references, and specialized protocols across multiple medical disciplines. Content is partitioned by clinical domain and indexed for both sparse (BM25) and dense (1,536-dimension embeddings with HNSW indexing) retrieval.

3.2 Query Dataset

500 clinical queries were generated across 6 specialties (Table 1) and 2 complexity levels. Queries were designed to represent the natural distribution of clinical information needs in Brazilian medical practice: pharmacology (drug interactions, dosing), clinical decision-making (differential diagnosis, management), emergency protocols, administrative coding, pediatric-specific queries, and guideline lookup.

Table 1: Query distribution by specialty and complexity.

Specialty	Simple	Complex	Total
Pharmacology	97	—	97
Clinical Decision	40	60	100
Emergency	57	23	80
Pricing/Coverage	75	—	75
Pediatrics	48	22	70
Clinical Protocols	45	25	70
Total	**362**	**130**	**500**

3.3 Retrieval Configurations

Five configurations were evaluated against all 500 queries:

1. **BM25-only:** Sparse keyword retrieval with medical synonym expansion.
2. **Dense-only (t=0.20):** Semantic retrieval at default permissive threshold.
3. **Dense-only (t=0.60):** Semantic retrieval at calibrated strict threshold.
4. **Hybrid, no authority:** BM25 + dense merged via RRF (k=60), no source weighting.
5. **Hybrid + authority:** RRF fusion with multiplicative boosts favoring national protocols over primary studies. Sources are classified into five tiers; exact weight values are calibrated per deployment.

3.4 Relevance Annotation

Two independent annotators labeled 1,487 (query, source) pairs from the hybrid + authority configuration. Annotation criterion: a source is relevant (1) if it directly addresses the drug, condition, or protocol in the query; otherwise not relevant (0). Inter-annotator agreement was measured by Cohen's κ [14]. Precision@5 was computed using annotator 1 labels.

3.5 Citation Verification

Two citation methods were compared across all 500 queries. Prompt-based: the LLM is instructed to generate inline citations from provided context. Deterministic: citations are injected programmatically by matching generated claims to retrieved document identifiers via exact matching. A citation is counted as correct if it references a document present in the retrieved set.

3.6 Statistical Tests

McNemar's test [15] assessed whether BM25 and hybrid retrieval surface the same documents (H_0: both methods retrieve identical sets). Fisher's exact test assessed whether deterministic verification is superior to prompt-based citation. Cohen's κ measured inter-annotator agreement. Significance threshold: α=0.05.

4 Results

4.1 RQ1: BM25 and Hybrid Retrieve Distinct Document Sets

Table 2: Source recall by retrieval configuration (n=500).

Configuration	Recall	Avg Rel. Score
BM25-only	99.0% (495/500)	0.232
Dense-only (t=0.20)	77.8% (389/500)	0.420
Dense-only (t=0.60)	20.2% (101/500)	0.752
Hybrid, no authority	95.6% (478/500)	0.340
Hybrid + authority	96.6% (483/500)	0.342

McNemar's test on the union of unique (query, source) pairs retrieved by BM25 and hybrid:

- A = 385 (retrieved by both methods)
- B = 467 (BM25-only)
- C = 680 (hybrid-only)

$$\chi^2 = \frac{(|467 - 680| - 1)^2}{467 + 680} = 39.55, \quad p < 0.001 \qquad (1)$$

Finding 1: BM25 and hybrid retrieval are statistically complementary ($p < 0.001$). They are not interchangeable. BM25 surfaces 467 unique sources unavailable to hybrid; hybrid surfaces 680 unique sources unavailable to BM25. Maximum corpus coverage requires both retrieval paths.

Specialty-level analysis reveals where complementarity is strongest:

Table 3: Hybrid recall and delta vs. BM25 by specialty.

Specialty	Compl.	n	Recall	Δ BM25
Emergency	complex	23	100%	0.00
Pediatrics	complex	22	100%	0.00
Clinical decision	complex	60	97%	+0.05
Pharmacology	simple	96	97%	−0.09
Clinical protocols	complex	25	92%	−0.24
Clinical protocols	simple	45	91%	−0.27

Clinical protocol queries show the largest BM25 advantage (Δ=−0.24 to −0.27), consistent with high term overlap between standardized protocol terminology and clinical queries. Clinical decision queries show a hybrid advantage (Δ=+0.05), consistent with greater semantic ambiguity in diagnostic reasoning.

4.2 RQ2: Dense-Only Retrieval Is Structurally Unreliable

Dense-only retrieval at default threshold (t=0.20) fails for 22.2% of queries (111/500). At calibrated threshold (t=0.60), failure rate increases to 79.8% (399/500).

Finding 2: At no threshold does dense-only retrieval achieve acceptable recall for clinical use. This is a corpus characteristic: standardized Brazilian clinical terminology produces high keyword overlap, making BM25 inherently strong and dense retrieval unnecessary for the majority of queries. Dense retrieval adds value only through complementary coverage (Finding 1), not as a replacement.

4.3 RQ3: Authority Weighting Affects Ranking, Not Recall

Hybrid retrieval with and without authority weighting achieved 96.6% and 95.6% recall respectively. The 1.0 percentage point difference (5 queries) is not statistically significant.

Table 4: Precision@5 by specialty (hybrid + authority).

Specialty	P@5	n
Clinical protocols	68.8%	64
Clinical decision	38.2%	96
Pharmacology	23.1%	97
Emergency	19.4%	79
Pricing/coverage	17.9%	78
Pediatrics	16.4%	69
Overall	**29.8%**	**483**

Overall NDCG@5 = 0.354. The wide range across specialties (16.4%–68.8%) reflects corpus coverage rather than retrieval algorithm performance.

Finding 3: Authority weighting is a ranking intervention, not a recall intervention. Its effect is measurable only through relevance ordering within the top-k results, not through coverage metrics. This finding has practical implications: teams evaluating authority-weighting systems should measure NDCG and human-judged ranking quality, not recall.

4.4 RQ4: LLM-as-Judge Is Valid for Portuguese Clinical Text

1,487 (query, source) pairs were independently annotated by two LLM judges.

Table 5: Inter-annotator agreement.

Metric	Value
Annotator 1 relevant	460/1,487 (30.9%)
Annotator 2 relevant	488/1,487 (32.8%)
Pairwise agreement	1,457/1,487 (98.0%)
Cohen's κ	0.954
Interpretation	Almost perfect [14]

Finding 4: LLM-based relevance annotation achieves κ=0.954 for Portuguese clinical text, meeting the "almost perfect agreement" threshold ($\kappa > 0.80$). This validates

LLM-as-judge methodology [11] for a non-English clinical domain, enabling scalable retrieval evaluation without requiring physician annotation time for every assessment. To our knowledge, this is the first reported κ for LLM-based clinical relevance judgment in Portuguese.

4.5 Citation Verification

Table 6: Citation verification (n=500).

Method	Correct	Wrong	Accuracy
Prompt-based	1/500	499/500	0.2%
Deterministic	461/500	39/500	92.2%

Fisher's exact $p < 0.001$, OR $\gg 1$. The 39 deterministic "failures" are queries where zero sources were retrieved—the system correctly returns no citations rather than fabricating them. Among queries with retrieved sources, the structural hallucination rate is 0%.

5 Discussion

5.1 Implications for Clinical RAG in Portuguese

Our results challenge three common assumptions in clinical RAG design.

First, the assumption that dense retrieval dominates BM25 for medical queries does not hold for Portuguese clinical text. BM25 achieves 99.0% recall—higher than any dense configuration—because standardized Brazilian clinical terminology (protocol names, drug names, diagnostic codes) produces high lexical overlap between queries and documents. This contrasts with findings for English [3], where paraphrasing and synonym variation make dense retrieval essential.

Second, hybrid retrieval is not redundant with BM25. Despite BM25's strong recall, hybrid surfaces 680 documents invisible to keyword matching. These tend to be semantically related content using different terminology—critical for complex clinical decision queries where physicians describe symptoms rather than name protocols.

Third, authority weighting—boosting national guidelines over primary studies—is commonly assumed to improve overall retrieval quality. Our evaluation shows it affects ranking within the result set but does not improve recall. This means authority weighting should be evaluated using ranking-sensitive metrics (NDCG, expert-judged top-k quality), not coverage metrics.

5.2 Implications for Retrieval Evaluation

The κ=0.954 finding for LLM-as-judge in Portuguese clinical text has immediate practical value. Physician annotation is expensive and rate-limiting for retrieval evaluation. If automated judgments achieve near-perfect agreement, teams can evaluate retrieval changes continuously without blocking on physician availability. This is particularly relevant for clinical AI systems in languages where annotator pools are smaller than in English.

5.3 Limitations

The 500-query dataset was synthetically generated to ensure specialty coverage. Real physician query distributions may differ. Precision@5 of 29.8% reflects corpus coverage gaps, not retrieval algorithm limitations—expanding the corpus would improve this metric without algorithmic changes. The LLM-as-judge validation, while strong (κ=0.954), should be confirmed against physician annotations. Findings are specific to Brazilian Portuguese clinical text and may not generalize to other languages without replication.

5.4 Future Work

Three directions follow from these findings. First, evaluation against real physician queries (currently encrypted per LGPD) would validate whether synthetic query distributions match production patterns. Second, a formal comparison of authority weighting using physician-judged ranking quality (rather than automated metrics) would quantify its clinical value. Third, replication of the BM25-hybrid complementarity finding for Spanish and French clinical text would establish whether it generalizes across Romance languages or is specific to Portuguese medical terminology.

6 Ethics Statement

This work evaluates a clinical decision support system, not a diagnostic tool. The system communicates that it assists but does not replace clinical judgment. All evaluation data was de-identified before analysis. No patient-identifiable information was used. The system complies with LGPD for sensitive health data.

7 Conclusion

We presented an empirical evaluation of retrieval strategies for Portuguese clinical text across 500 queries and 6 specialties. Four findings contribute to the understanding of RAG in non-English clinical domains:

1. BM25 and hybrid retrieval are statistically complementary ($p < 0.001$)—neither approach is sufficient alone for maximum clinical corpus coverage.
2. Dense-only retrieval fails 22.2% of Portuguese clinical queries at default threshold, making it structurally unsuitable as the sole retrieval path for this domain.
3. Authority weighting affects ranking, not recall—evaluation methodology must match the intervention type.

4. LLM-as-judge achieves κ=0.954 for Portuguese clinical relevance, validating automated evaluation for this domain.

Additionally, deterministic citation verification eliminates hallucinated references entirely (461/500 vs. 1/500 for prompt-based, Fisher $p < 0.001$).

These findings apply to any team building clinical RAG systems for Portuguese—and potentially for other non-English clinical domains where standardized terminology produces high keyword overlap. The evaluation methodology, including query dataset and annotation framework, is available as open-source software at https://github.com/nomad-link-id.

References

[1] P. Lewis, E. Perez, A. Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 33:9459–9474, 2020.

[2] C. Zakka, R. Chaurasia, A. Asber, et al. Almanac: Retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2), 2024.

[3] V. Karpukhin, B. Oğuz, S. Min, et al. Dense passage retrieval for open-domain question answering. *EMNLP*, pages 6769–6781, 2020.

[4] Z. Ji, N. Lee, R. Frieske, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[5] X. Ma, K. Sun, R. Pradeep, J. Lin. Zero-shot listwise document reranking with a large language model. *arXiv:2305.02156*, 2023.

[6] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *NeurIPS Datasets and Benchmarks*, 2021.

[7] G. Cormack, C. Clarke, S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *SIGIR*, pages 758–759, 2009.

[8] J. Chen, R. Xiao, H. Wu. Hybrid retrieval with reciprocal rank fusion for domain-specific QA. *ACL*, 2024.

[9] T. Gao, H. Yen, J. Yu, D. Chen. Enabling large language models to generate text with citations. *EMNLP*, 2023.

[10] S. Min, K. Krishna, X. Lyu, et al. FActScore: Fine-grained atomic evaluation of factual precision. *EMNLP*, 2023.

[11] L. Zheng, W. Chiang, Y. Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *NeurIPS*, 2023.

[12] P. Thomas, S. Spielman, N. Craswell, B. Mitra. Large language models can accurately predict searcher preferences. *SIGIR*, 2024.

[13] G. Guyatt, A. Oxman, G. Vist, et al. GRADE: An emerging consensus on rating quality of evidence. *BMJ*, 336:924–926, 2008.

[14] J.R. Landis, G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[15] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[16] K. Singhal, S. Azizi, T. Tu, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023.

[17] H. Nori, N. King, S. McKinney, D. Carignan, E. Horvitz. Capabilities of GPT-4 on medical competency examinations. *arXiv:2303.13375*, 2023.

[18] G. Xiong, Q. Jin, Z. Lu, A. Zhang. Benchmarking retrieval-augmented generation for medicine. *ACL Findings*, 2024.