Evaluating the Retrieval Component in a Retrieval-Augmented Summarization System for Patient Records in French

Marco Naguib[1,2], **Christel Gérardin**[3,4], **Victor Beaucoté**[5,6], **Cyril Charron**[5], **Adrien Joseph**[5,7], **Aurélie Névéol**[1], **Xavier Tannier**[3]

[1] Université Paris-Saclay, CNRS, LISN, Orsay, France
[2] Innovation and Data Unit, IT Department, AP–HP, Paris, France
[3] LIMICS, Université Sorbonne Paris-Nord, Inserm, Sorbonne Université, Paris, France
[4] Service de médecine interne, Hôpital Tenon, AP–HP, Paris, France
[5] Medical Intensive Care Unit, Ambroise Paré Hospital, AP–HP, Boulogne-Billancourt, France
[6] UFR Simone-Veil Santé, UVSQ, Montigny-le-Bretonneux, France
[7] Inserm, CESP, Paris-Saclay University, UVSQ, Villejuif, France

Abstract

In emergency and intensive care settings, clinicians must process large volumes of patient data to make time-sensitive decisions. Summarizing patient records can help reduce cognitive load and improve decision-making, but the complexity and variability of clinical documentation create challenges. This study explores a Retrieval-Augmented Generation (RAG) approach, consisting of two phases: (1) retrieval of relevant clinical information, and (2) generation of a summary. This paper evaluates the retrieval component of RAG systems, focusing on its performance in clinical contexts. Using French clinical text, we assess retrieval models and propose an annotation-based querying method to improve accuracy and consistency in retrieving core clinical information. We use an annotated dataset from the AP–HP hospital to benchmark retrieval models tailored for French clinical records. The proposed annotation-based querying method is compared to traditional prompt-based approaches, demonstrating improved retrieval performance. The findings indicate that specialized retrieval techniques enhance the effectiveness of RAG systems in clinical settings, providing more accurate and relevant information for summarization. The study contributes to the development of clinical decision support tools by improving the retrieval process in RAG systems. The proposed methods offer a structured approach to summarizing patient records, which may help clinicians manage information more efficiently.
Keywords: Electronic Health Records, Retrieval-Augmented Generation (RAG), Information Retrieval

1. Introduction

In emergency and intensive care, clinicians must rapidly process extensive patient data for critical decisions. Efficient summarization of patient records can alleviate cognitive burdens and enhance decision-making by ensuring vital details such as patient allergies are accessible at the point of care.

Summarizing patient records is inherently challenging. Clinical data is complex, variable over time, and often inconsistently documented, which complicates the development of generalizable summarization systems: (Pivovarov and Elhadad, 2015; Keszthelyi et al., 2023). Additionally, traditional metrics like BLEU and ROUGE may not effectively assess the quality of summarization systems because they are designed to capture surface similarity with a reference summary rather than factual accuracy and relevance.

Recent advances in Retrieval-Augmented Generation (RAG) architectures offer promising solutions to these challenges by explicitly grounding generated summaries in relevant source documents. RAG combines aspects of extractive summarization—through retrieving relevant document passages—with abstractive methods by constructing coherent summaries from these retrieved passages. While RAG has shown potential in general-domain summarization tasks: (Lewis et al., 2020; Liu et al., 2024; Edge et al., 2025), applications to clinical summarization—particularly in languages other than English—remain underexplored: (Alkhalaf et al., 2024; Ji et al., 2024). Specifically, there is a notable lack of systematic evaluation of retrieval methods tailored to clinical texts in French, as well as limited exploration of effective querying strategies that leverage clinical annotations.

In this study, we address these gaps by systematically evaluating and improving the retrieval component of RAG architectures for patient record summarization in French clinical settings. We focus on enhancing retrieval effectiveness, leaving the detailed evaluation of the generation component for future work.

The main contributions of this study include:

1. **Benchmarking Retrieval Methods for French Clinical Text:** We establish a systematic evaluation framework and provide comprehensive benchmarks comparing sparse (e.g., TF-IDF) and dense retrieval methods (sentence embedding approaches)

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 57–65
16 May 2026.

on French clinical documents, addressing an evaluation gap in the literature.

2. **Annotated Clinical Dataset:** We annotate real-world clinical documents provided by the AP–HP. This benchmark enables rigorous evaluation of retrieval and summarization techniques in complex clinical settings.

3. **Novel Annotation-Based Querying Method:** We propose an innovative querying approach that leverages clinical annotations, consistently outperforming traditional prompting methods and significantly enhancing retrieval performance, especially for computationally efficient methods like TF-IDF.

By establishing a robust evaluation framework and demonstrating the effectiveness of annotation-driven retrieval, our work lays a strong foundation for future research on generating structured, clinically relevant summaries that can effectively support medical decision-making. We share the code of our experiments at https://github.com/marconaguib/rag.

2. Related Work

Patient record summarization is essential in medical informatics for condensing extensive clinical texts into concise summaries that aid healthcare professionals: (Alkhalaf et al., 2024; Ji et al., 2024). Early methods relied on extractive techniques, selecting key sentences to maintain factual accuracy but often resulting in incoherent summaries: (Mishra et al., 2014; Moen et al., 2016).

Advancements in natural language processing have shifted focus to abstractive summarization, which generates novel text for more fluent and coherent summaries: (Zhang et al., 2018, 2020). However, ensuring factual accuracy and handling specialized medical terminology remain significant challenges: (Kryściński et al., 2020; Maynez et al., 2020).

The Retrieval-Augmented Generation (RAG) framework has shown promise in improving factual accuracy by grounding generated content in relevant source documents (Lewis et al., 2020). In clinical settings, RAG enhances both the accuracy and relevance of summaries by leveraging patient data: (Alkhalaf et al., 2024; Uapadhyay and Viviani, 2025).

Evaluating the factual correctness of summaries is critical, as traditional metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are insufficient for clinical accuracy (Zhang et al., 2020). Specialized evaluation frameworks such as FRAMES have been developed to better assess retrieval and reasoning in RAG systems (Krishna et al., 2024). Additionally, methods like Facts as a Function (FaaF) and factuality optimization techniques using reinforcement learning have been proposed to mitigate inaccuracies in generated text: (Katranidis and Barany, 2024; Cai et al., 2024).

Recent work has also explored explanable evaluation paradigms for summarization quality beyond aggregate scores. Herserant and Guigue (2025) propose a framework that decomposes summary evaluation into atomic sentence-level assessments, offering improved interpretability and detailed alignment between source text and generated summaries, addressing a key limitation of conventional evaluation metrics that only provide holistic scores.

Our work builds on these developments by applying RAG-based patient record summarization to real-world clinical data, with a particular emphasis on retrieval component effectiveness and innovative annotation-based querying methods. Additionally, we contribute the first systematic benchmark comparing sentence embeddings specifically tailored to French clinical documents, complementing recent general-domain evaluation frameworks (Ciancone et al., 2024), and further informing model selection and retrieval strategies in clinical NLP.

3. Data and Annotation

Our study leverages a corpus of 10,000 clinical documents from the AP–HP hospital, approved by an internal ethics review board. For our experiments, we selected individuals with a median number of documents per patient (=61) to represent typical clinical scenarios realistically.

The clinical document corpus was annotated by four critical care physicians. Annotators were instructed to identify and label clinically relevant information within patient records, focusing on essential details required for structured clinical summaries. To quantify annotation consistency and reliability, three patients were sampled for inter-annotator agreement computation. Additionally, eight other patients were independently annotated by a single annotator to expand our dataset, totaling 11 patients and 671 documents.

Annotations were categorized into six predefined clinical labels, selected to comprehensively represent patient profiles and effectively support clinical decision-making:

Lifestyle: Lifestyle-related details such as living arrangements and daily activities that can impact health outcomes.

Medical history: Key elements of the broader medical history, including chronic disorders which may impact treatment options.

Surgical history: Details regarding the patients surgical history, providing context for current health status.

Treatments: Data on medications and therapies received, which are essential for understanding ongoing care.

Allergies: Information about the patients allergies is crucial for avoiding adverse effects.

Biometrics: Critical biometric measurements (e.g., weight, height) that inform clinical decisions.

The category definitions for clinical features were used as annotation guidelines.

4. Experiments

4.1. Retrieval-Augmented Generation Pipeline

Our Retrieval-Augmented Generation (RAG) pipeline is illustrated in Figure 1. It consists of two sequential steps—retrieval and generation—applied iteratively to construct a structured clinical summary from heterogeneous patient records.

Step 1: Retrieval Clinical documents are first segmented into fixed-size text chunks, which constitute the retrieval units. Then, for each target clinical section (e.g., medical history), relevant passages are retrieved from the patient record using either vocabulary-based methods (TF-IDF, BM25) or transformer-based embedding models. As shown in Figure 1, this step acts as a filtering mechanism that selects a small set of section-specific excerpts from a large and unstructured collection of clinical documents.

Step 2: Generation In the second step, a large language model (LLM) synthesizes the retrieved passages into a coherent and structured summary tailored to the clinical section of interest. Rather than processing the entire patient record, the LLM operates on the targeted subset produced by the retrieval step, which constrains the input and guides generation.

Although generation is an essential component of the overall pipeline, the analysis presented in this work focuses exclusively on Step 1, namely retrieval performance. Preliminary experiments conducted with the complete pipeline on our clinical data (to be reported in the supplementary materials in the final version) indicate that effective retrieval remains critical, even when using large-context LLMs. In particular, when provided with entire patient records instead of retrieved excerpts, the model tended to generate overly long and poorly structured summaries, whereas retrieval-based inputs led to more focused and clinically relevant outputs.

4.2. Retrieval Queries

We investigated seven query formulation strategies aimed at improving retrieval performance and, in turn, the quality of patient record summarization. These strategies include both manually defined textual queries and data-driven queries automatically derived from annotated clinical data.

Textual Queries. The first five strategies rely on explicit textual formulations expressed in natural language. They differ in how the target information is described and constrained, but all provide a text-based query to the retriever:

- **Specification-Based Prompts**: queries derived from clinician-defined specifications that explicitly describe the expected content of each summary section, closely mirroring clinical guidelines.
- **Examples**: queries composed of representative example phrases illustrating typical linguistic realizations of the targeted clinical information.
- **Keywords**: curated lists of salient terms and entities associated with each clinical section, enabling focused lexical or semantic matching.
- **Contextual Prompts**: queries that situate the requested information within its broader clinical context, encouraging retrieval of passages jointly mentioning related concepts, temporal cues, or recent assessments.
- **Structure Prompts**: queries that explicitly specify the expected organizational structure by enumerating fields or subcomponents to be retrieved.

Data-Driven Queries. The remaining two strategies construct query representations automatically from annotated clinical data, without relying on manually written text queries. Their underlying principle is illustrated in Figure 2.

- **Query-from-Annotations**: for a given clinical label, all annotated spans are encoded and averaged to form a single query vector that serves as a semantic prototype of the target concept.
- **Query-from-Annotation-Chunks**: a chunk-level variant in which all text chunks containing at least one annotation of the target label are encoded and averaged, capturing a broader contextual signal (Figure 2).

The two data-driven strategies are evaluated in a leave-one-out cross-validation (LOOCV) setting. For each experiment, query vectors evaluated on

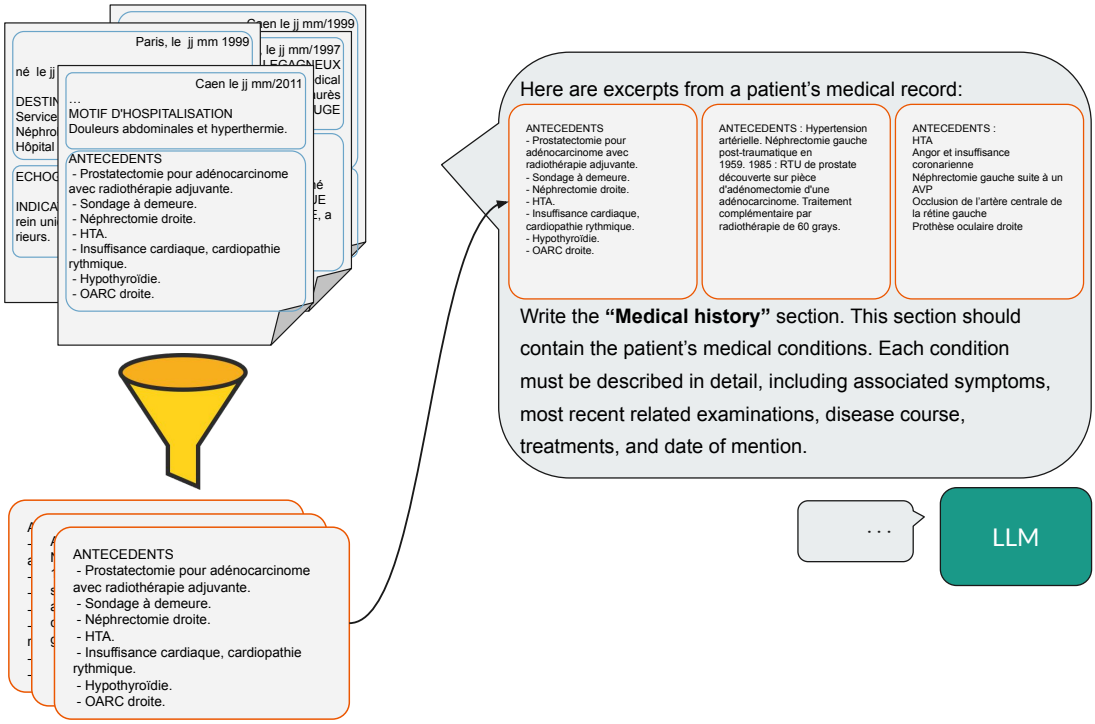

Figure 1: Overview of the Retrieval-Augmented Generation (RAG) pipeline for structured clinical summarization. Relevant passages are first retrieved from heterogeneous patient records for a given clinical section (e.g., medical history) and then provided to a large language model to generate a section-specific summary.

one patient record are constructed exclusively from annotations belonging to other patient records, ensuring that annotations from the test patient record are never used during query construction.

To evaluate the effectiveness of these query strategies, we benchmark them using a range of sparse and dense retrieval models, described next.

4.3. Retrieval Models

We evaluate seven retrieval models, selecting the top-performing open-source sentence embedding models for French as identified by the recent general-domain benchmark of Ciancone et al. (2024). Our evaluation complements this benchmark by systematically comparing these embeddings specifically on clinical texts.

TF-IDF: Traditional retrieval method based on term frequency-inverse document frequency.

BM25: Probabilistic retrieval method extending TF-IDF with term saturation and document length normalization. Note that BM25 is incompatible with our annotation-based querying approach, as it relies on explicit query terms rather than standalone query embeddings derived from annotations.

dangvantuan/sentence-camembert-large: CamemBERT-based sentence transformer optimized for French semantic retrieval.

BAAI/bge-m3: Multilingual embedding model from BAAI, optimized for diverse retrieval tasks.

OrdalieTech/Solon-embeddings-large-0.1: Large-scale embedding model designed for specialized-domain retrieval.

manu/sentence_croissant_alpha_v0.4: French embedding model emphasizing semantic similarity.

jinaai/jina-embeddings-v3: Multilingual embeddings optimized for efficient neural search.

4.4. Metrics

We evaluate the retrieval performance using recall at various reduction rates, specifically retrieving the top 5%, 10%, 15%, and 20% of chunks within each patient profile (i.e., the passage-level retrieval units described in Section 4.1). For each reduction rate, we compute the recall across all annotated labels to assess how effectively each retrieval model retrieves the most pertinent information.

In addition, we measure the computational efficiency of each model by recording the average computation time on an NVIDIA A100 GPU and on a 64-core CPU. This dual assessment helps us understand the trade-off between retrieval accuracy and computational demands, which is critical for practical deployment in clinical environments.

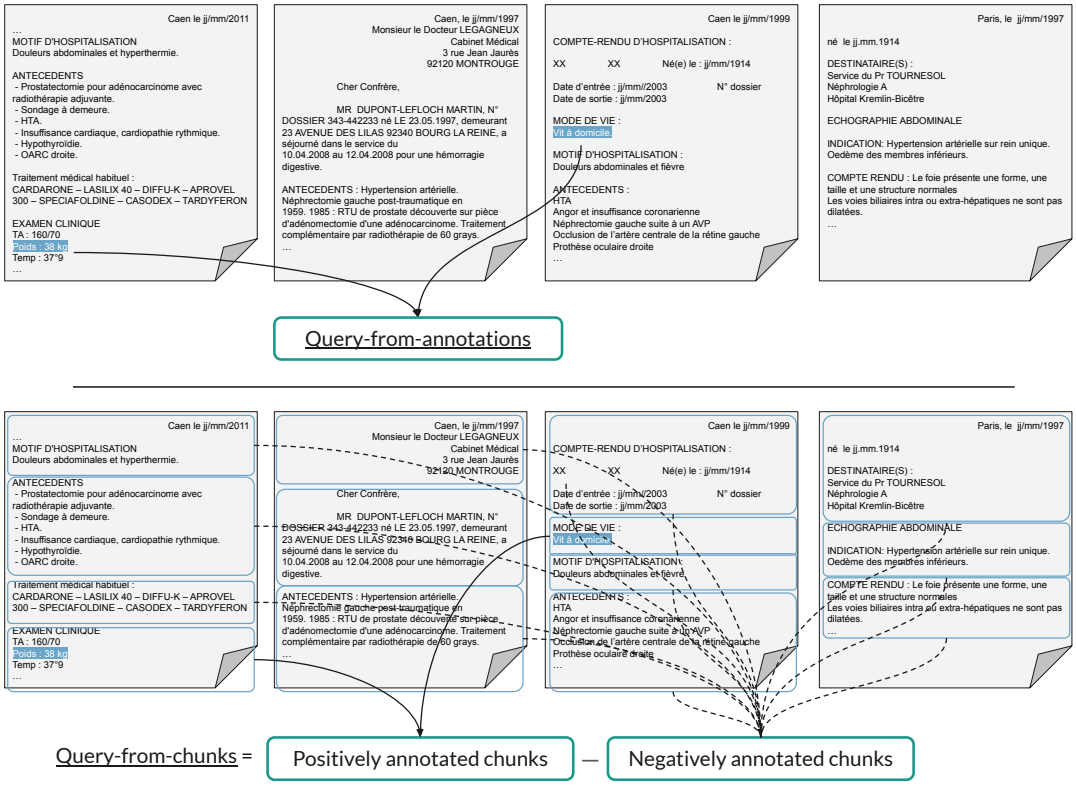

Figure 2: Illustration of data-driven query construction from annotations. Annotated spans or chunks associated with a given clinical label are extracted from multiple patient records and encoded into vector representations. In the *Query-from-Annotations* approach, only annotated spans are averaged to form a query vector, whereas *Query-from-Annotation-Chunks* averages representations of all text chunks containing at least one annotation. The resulting query vectors are then used to retrieve relevant passages in unseen patient records.

5. Results and Discussion

5.1. Inter-annotator Agreement

Inter-annotator agreement was measured to assess annotation consistency across clinical categories (Table 1). Overall agreement was moderate (average Cohen's $\kappa = 0.56$), reflecting both the intrinsic complexity of clinical records and subtle differences in annotation practices. Agreement varied across categories, with higher consistency for Lifestyle, Surgical History, and Allergies, and lower agreement for Medical History and Biometrics.

Pairwise analyses further highlight this variability, with κ values ranging from 0.43 to 0.86 depending on the entity type and annotator pair. Categories such as Lifestyle and Allergies exhibit consistently high agreement, whereas Medical History and Biometrics show greater divergence, underscoring the ambiguity of these clinical concepts.

To account for this variability, we adopted an additive consensus approach for the first three patients, aggregating annotations from all annotators when evaluating retrieval performance. Recall was computed against this merged reference, ensuring that the evaluation captures the full spectrum of expert judgment and mitigates the impact of individual annotation differences. This strategy provides a more robust and comprehensive benchmark for assessing retrieval and summarization performance under realistic clinical uncertainty.

Table 1: Inter-annotator agreement by category. Pairwise κ reported as min–max.

Category	Avg. κ	Median κ	Min–Max
Overall	**0.56**	**0.56**	0.51–0.63
Lifestyle	0.79	0.80	0.72–0.86
Medical history	0.52	0.51	0.45–0.62
Surgical history	0.67	0.67	0.61–0.73
Treatments	0.60	0.60	0.57–0.63
Allergies	0.66	0.66	0.49–0.81
Biometrics	0.52	0.52	0.43–0.66

5.2. Retrieval

Figure 3 shows the macro-average recall at 20% reduction for all query types and models. Using annotations as queries consistently improves retrieval performance compared to hard prompts. Notably, TF-IDF benefits significantly from annotation-based

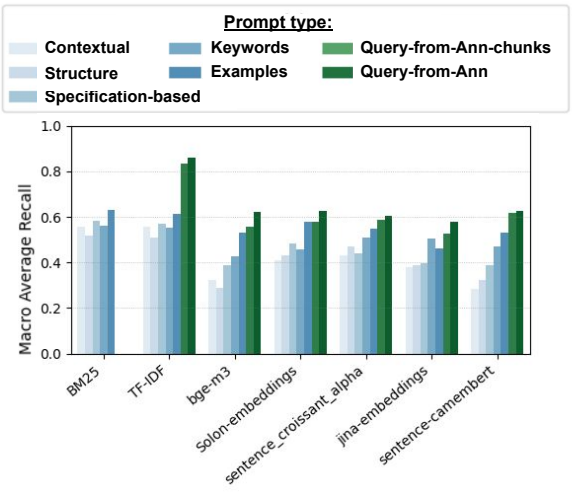

Figure 3: Macro-average Recall at 20% reduction per retrieval model and query type. Annotation-based queries (green) consistently improve retrieval performance.

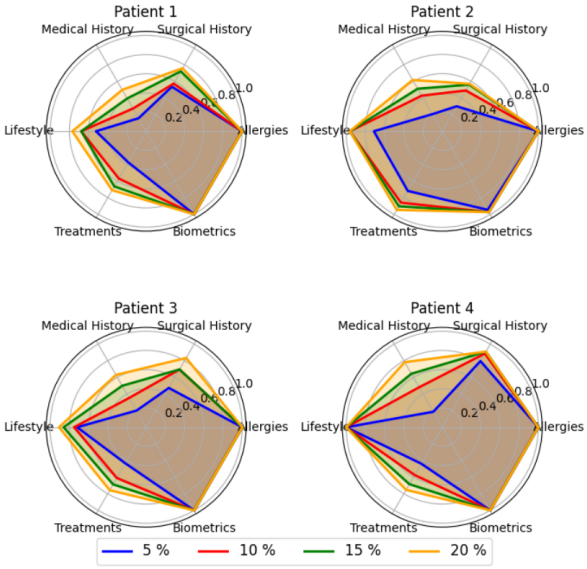

Figure 4: Recall across clinical sections for individual patient profiles at different reduction rates.

querying, surpassing transformer-based models in recall and being about 100 times faster when a GPU is available, while also being CPU-friendly (Table 2). Given its computational efficiency and strong retrieval performance, TF-IDF is particularly suitable for clinical settings, where speed and reliability are critical.

Finally, Figure 4 illustrates retrieval recall across different annotation categories for four patient profiles, highlighting variability in retrieval difficulty across clinical information types. These profiles include the three patients evaluated using the Additive consensus approach (1, 2, and 3). As anticipated by clinicians, performance is higher for

Table 2: Average computation times of encoding all patient profile chunks on an NVIDIA A100 GPU and on a 64-core CPU.

Model	NVIDIA A100 (s)	64-core CPU (s)
sentence-camembert	4.88	160
bge-m3	4.76	170
Solon-embeddings	4.71	173
sentence-croissant-alpha	19.4	716
jina-embeddings	2.18	161
TF-IDF	-	0.05
BM25	-	0.03

compact categories such as *allergies* compared to *medical history*, where information is scattered across documents and changes more over time.

6. Conclusion

This work investigated the retrieval component of Retrieval-Augmented Generation (RAG) systems for patient record summarization in French clinical settings. Through a systematic evaluation of seven retrieval models and multiple query formulation strategies, we demonstrated that retrieval effectiveness can be substantially improved by leveraging clinical annotations to construct query representations. Across models and reduction rates, annotation-based queries consistently outperformed manually designed textual prompts, highlighting the importance of data-driven querying strategies for clinical information retrieval.

A key finding of our study is that simple and computationally efficient methods, such as TF-IDF, can achieve strong retrieval performance when coupled with annotation-derived queries—often matching or surpassing transformer-based embedding models at a fraction of the computational cost. This result is particularly relevant for real-world clinical deployment, where robustness, interpretability, and resource efficiency are critical constraints. Our benchmark further provides the first systematic comparison of sentence embedding models on French clinical documents, complementing existing general-domain evaluations with clinically grounded insights.

By isolating and rigorously evaluating the retrieval stage, this work establishes a solid foundation for the development and assessment of retrieval-augmented clinical summarization systems. Future work will extend this evaluation to the generation component, with a focus on factual accuracy, clinical relevance, and end-user utility, as well as on reducing reliance on manual annotations to improve scalability.

7. Acknowledgments

This work has received funding from the PARTAGES project, awardee of the Bpifrance France 2030 call for proposals “Digital Commons for Generative Artificial Intelligence.”

Limitations

Our evaluation primarily focuses on the retrieval step of the proposed retrieval-augmented summarization pipeline. While retrieval quality is a critical component, it does not necessarily translate directly into the overall performance of the full system, particularly in terms of summarization accuracy and clinical usability. Future work will aim to assess the impact of retrieval effectiveness on downstream tasks, such as information synthesis and generation, to provide a more holistic evaluation of the pipeline.

Furthermore, our evaluation concentrated on retrieval performance and computational efficiency but did not comprehensively address other important factors such as potential biases in retrieval outcomes or the interpretability of retrieved results. These aspects remain open areas for further investigation, especially given the clinical context in which fairness, transparency, and robustness are paramount.

The annotated corpus used in this study also presents certain limitations. First, due to confidentiality agreements, we are unable to publicly share the dataset or the corresponding annotations, restricting reproducibility and external validation. While this is common in clinical research, it underscores the broader challenge of benchmarking patient record summarization models on real-world data. Second, the number of patient records in the corpus was relatively small, reflecting the pilot nature of this study. Despite this, the annotations demonstrated moderate to substantial agreement, suggesting that expert consensus was reasonably achieved. The annotation process involved 671 documents, highlighting the complexity and effort required in manual Electronic Health Record (EHR) summarization.

Additionally, the current methodology relies heavily on high-quality clinical annotations to construct effective query representations. While this allows for precise and structured retrieval, it also limits scalability and generalizability, as such annotations are time-consuming and costly to obtain in broader settings. Exploring strategies for reducing annotation dependence—such as semi-supervised methods or automated labeling—will be an important direction for future work.

Future work will seek to address these limitations by scaling the dataset, refining evaluation metrics, and incorporating a broader range of factors—including potential biases and real-world clinical applicability—into our assessment framework.

Ethical Considerations

This study involves the use of de-identified patient data approved (CSE 21-36) by an internal ethics review board at the AP–HP hospital. Throughout our work, we made every effort to maintain patient privacy and confidentiality in compliance with ethical and institutional guidelines.

The application of recommendation or automated summarization systems in clinical care carries potential risks, such as automation bias, where clinicians may overly trust system outputs without sufficient scrutiny. Such biases can substantially impact patient outcomes if incorrect or incomplete information goes unnoticed.

Our findings specifically highlight the value of simpler and more interpretable methods, such as TF-IDF retrieval, which not only match but can outperform sophisticated sentence embedding models when coupled with our query-from-annotation method. Besides offering improved interpretability, TF-IDF is computationally lighter, thus aligning better with environmentally responsible and sustainable AI practices.

We recognize the critical implications of integrating automated tools into clinical workflows. Therefore, we emphasize the necessity of extensive evaluation beyond retrieval performance. Our future research steps include a comprehensive assessment of the full retrieval-augmented generation (RAG) pipeline, addressing not only factual accuracy and summary quality but also clinical relevance, user trust, overall usability, and eventual impacts on clinical decisions and patient outcomes.

8. Bibliographical References

Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. 2024. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, 156:104662.

Tianchi Cai, Zhiwen Tan, Xierui Song, Tao Sun, Jiyan Jiang, Yunqi Xu, Yinger Zhang, and Jinjie Gu. 2024. FoRAG: Factuality-optimized Retrieval Augmented Generation for Web-enhanced Long-form Question Answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 199–210, New York, NY, USA. Association for Computing Machinery.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. MTEB-French: Resources for French Sentence Embedding Evaluation and Analysis. *arXiv preprint arXiv:2405.20468*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization.

Tanguy Herserant and Vincent Guigue. 2025. Seval-ex : Un paradigme basé sur les phrases atomiques pour une évaluation explicable de la qualité des résumés. In *Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA)*, pages 217–229, Marseille, France. ATALA \\& ARIA.

Yuelyu Ji, Zhuochun Li, Rui Meng, Sonish Sivarajkumar, Yanshan Wang, Zeshui Yu, Hui Ji, Yushui Han, Hanyu Zeng, and Daqing He. 2024. RAG-RLRC-LaySum at BioLaySumm: Integrating Retrieval-Augmented Generation and Readability Control for Layman Summarization of Biomedical Texts.

Vasileios Katranidis and Gabor Barany. 2024. FaaF: facts as a function for the evaluation of RAG systems. *arXiv e-prints*, pages arXiv–2403.

Daniel Keszthelyi, Christophe Gaudet-Blavignac, Mina Bjelogrlic, Christian Lovis, et al. 2023. Patient information summarization in clinical settings: scoping review. *JMIR Medical Informatics*, 11(1):e44639.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Shengjie Liu, Jing Wu, Jingyuan Bao, Wenyi Wang, Naira Hovakimyan, and Christopher G Healey. 2024. Towards a Robust Retrieval-Based Summarization System.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R. Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics*, 52:457–467. Special Section: Methods in Clinical Research Informatics.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Rishabh Uapadhyay and Marco Viviani. 2025. Enhancing Health Information Retrieval with RAG by Prioritizing Topical Relevance and Factual Accuracy. *arXiv preprint arXiv:2502.04666*.

Yuhao Zhang, Qian Chi, and Xibin Xia. 2018. Learning to Summarize Radiology Findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.