Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Cluster-Aware Retrieval-Augmented Generation (RAG) with Hybrid Retrieval for Faithful Medical Report Summarization

Kiarash Torabizadeh[†,*], Rachid Hedjam[†,◇], Mebarka Allaoui [†], Bessam Abdulrazak[‡]
[†] Department of Computer Science, Bishop's University, Sherbrooke, QC, Canada
[‡] Computer Science Department, University of Sherbrooke, Sherbrooke, QC, Canada
[◇] University of Quebec at Outaouais (UQO), Gatineau, QC, Canada

Abstract

Large Language Models (LLMs) can generate fluent medical summaries but may hallucinate facts not supported by source clinical text, limiting safe clinical adoption. In our prior work, we improved relevance through embedding-based patient clustering and cluster-wise GPT-4.0 summarization; however, summaries could still include unsupported claims due to a lack of explicit evidence grounding. This paper extends that pipeline with a cluster-aware Retrieval-Augmented Generation (RAG) layer to ground summaries in retrieved evidence. For each cluster, we construct two evidence artifacts: (i) a Cluster Profile aggregating clinical statistics (e.g., means, ranges, abnormality rates), and (ii) a Snippet Bank of patient report excerpts. Evidence is retrieved via a hybrid retriever that combines TF-IDF and dense-embedding similarity with weighted scoring. We enforce citation-constrained prompting, requiring each major claim to cite retrieved evidence or be marked as "insufficient evidence". We evaluate cluster-wise RAG summaries using metrics for faithfulness (supported-claim rate), completeness (coverage of key abnormal indicators), and safety and overreach (diagnostic, medication, and absolute claims). Experiments on a synthetic hypertension dataset (150 patients stratified into low-, average-, and high-risk) show that our approach reduces hallucinations while preserving the personalization benefits of clustering. The source code of our framework is available on Github.

Keywords: Large Language Models (LLM), Retrieval-Augmented Generation (RAG), Manifold Embedding, Patient Clustering, Medical Report Summarization, Faithfulness, Hybrid Retrieval.

1. Introduction

Clinical documentation is central to modern healthcare, but it is also a major source of clinician workload. Electronic health records contain long, repetitive, and heterogeneous narratives that must be reviewed quickly and accurately. Automated medical summarization has therefore emerged as an assistive technology intended to reduce cognitive burden while preserving clinically relevant information [1]. Recent large language models (LLMs) have substantially improved the fluency of abstractive summaries [2]. However, in the medical domain, linguistic quality alone is insufficient: the primary barrier to adoption is *factual reliability*. LLMs may generate plausible but unsupported statements, exaggerate risk factors, or imply diagnoses not grounded in the record [3, 4]. Such hallucinations can mislead interpretation, reduce trust, and affect downstream decisions.

Our previous study addressed a related issue of *context mismatch*. Summarizing heterogeneous patient populations with a single model often yields overly generic descriptions. Patients were grouped according to clinical characteristics, and cluster-wise summaries were generated [5], improving relevance by conditioning summaries on coherent patient groups. An important component of this pipeline is the use of embedding-based representations before clustering. Clinical variables are inherently heterogeneous and may exhibit complex, non-linear relationships that are not well captured in the original feature space. By projecting patient data into a lower-dimensional embedding space, using techniques like PCA, t-SNE [6], UMAP [7], t-SNE-PSO [8], and SN-tSNE [9], latent clinical structures become more explicit, enabling the formation of more coherent and semantically meaningful patient groups. This step is critical, as the quality of clustering directly determines the relevance and consistency of the downstream summaries. Nevertheless, contextual relevance alone did not ensure correctness, since the model could still introduce unsupported claims. This limitation reflects a deeper problem: current LLM summarization relies largely on parametric knowledge rather than verifiable evidence. In clinical settings, summaries must be *auditable*, meaning each statement should be traceable to observable patient data. Relative to our prior work, the contribution of the present paper is not the clustering stage itself, but the introduction of an explicit evidence-construction and retrieval-grounding layer that constrains cluster-level generation and enables claim-level reliability analysis.

[*]ktorabizadeh23@ubishops.ca

Retrieval-Augmented Generation (RAG) addresses this issue by conditioning generation on retrieved evidence instead of model memory [10]. Grounded generation has been shown to reduce unsupported claims and provide citation-based provenance that helps users verify outputs [11]. However, existing RAG approaches operate at the document level and ignore population structure. Medical reports are not independent documents: they describe related patients sharing clinical patterns, and treating them as a flat corpus may retrieve irrelevant or contradictory evidence.

Evaluating such systems also requires metrics beyond textual similarity. Lexical measures such as ROUGE do not capture factual correctness [3]. Recent work, therefore, performs claim-level verification, decomposing summaries into atomic statements and checking each against supporting evidence [12]. This perspective emphasizes reliability and traceability rather than stylistic similarity. We therefore propose a *cluster-aware Retrieval-Augmented Generation framework* for medical report summarization. Patients are first organized into clinically coherent groups, and summaries are generated from a structured evidence space consisting of (i) aggregated clinical statistics and (ii) representative report excerpts. Each major statement must cite retrieved evidence, enforcing traceability and improving reliability. The contributions of this paper are summarized as follows:

- A cluster-aware RAG paradigm integrating population structure with evidence-grounded LLM summarization.
- A dual-evidence representation consisting of a structured *Cluster Profile* and a textual *Snippet Bank* for verifiable generation.
- A hybrid retrieval mechanism combining lexical and semantic similarity to select clinically relevant evidence.
- An evaluation protocol focused on clinical reliability, including faithfulness, completeness, and safety/overreach analysis beyond traditional similarity metrics.

2. Proposed Method

Our framework generates clinically meaningful summaries that are both context-aware and evidence-grounded (Fig. 1). The central idea is to separate three roles that are often conflated in LLM summarization: (i) organizing heterogeneous patient records into coherent clinical populations, (ii) constructing verifiable evidence describing each population, and (iii) constraining language generation so that statements are supported by that evidence. Instead of summarizing raw reports directly, the model summarizes structured and retrieved evidence derived from clinically similar patients. In this paper, the main extension beyond our prior clustering-based pipeline is the addition of the evidence-construction, retrieval, and citation-constrained generation stages.

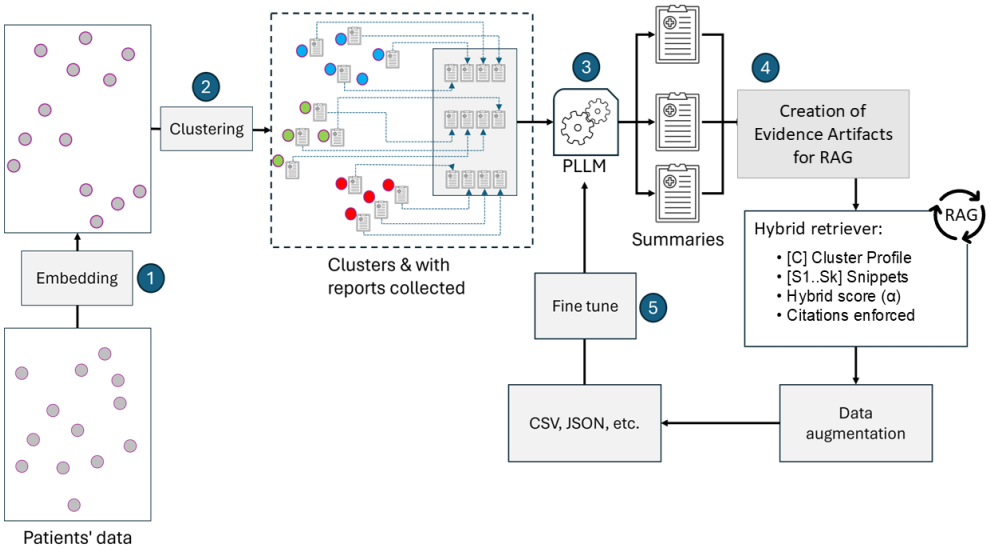

Figure 1. Overview of the proposed cluster-aware RAG framework for medical report summarization. Steps 1–3: Organize patients into clinically similar groups using embedded structured variables and clustering. Step 4 constructs the cluster-level evidence space, consisting of a Cluster Profile and a Snippet Bank, and retrieves the most relevant evidence using hybrid retrieval. Step 5 generates a grounded cluster-level summary constrained by the retrieved evidence.

Population structuring via embedding and clustering (Steps 1–3). Medical reports originate from patients with heterogeneous characteristics, and summarizing all reports jointly often yields generic descriptions. We therefore organize patients into clinically coherent populations. Structured clinical variables are projected into a lower-dimensional representation using the embedding procedure of [5], capturing relationships among indicators such as blood pressure, BMI, glucose, and cholesterol. K-Means clustering partitions the embedded data into distinct clinical profiles. For each cluster, associated reports are aggregated, and an initial cluster-level narrative is generated by a large language model. Because this stage relies only on loosely structured text and model priors, unsupported generalizations may still occur; subsequent stages introduce explicit grounding.

Evidence construction (Step 4A). For each cluster, we construct an explicit evidence space composed of two complementary artifacts. *Cluster Profile (structured evidence):* a statistical summary including sample size, mean, and range of key indicators, and abnormality prevalence rates (e.g., hypertension, obesity, elevated cholesterol), providing quantitative grounding. *Snippet Bank (textual evidence):* representative report excerpts annotated with cluster and patient identifiers, providing natural-language justification for observations not captured by statistics. Together, structured and textual evidence support both population-level and narrative claims. This step serves as evidence enrichment for each cluster by transforming raw cluster contents into a compact evidence space used later for retrieval-grounded generation.

Hybrid evidence retrieval (Step 4B). During generation, the model retrieves the most relevant evidence within the cluster rather than using all available data. We combine lexical and semantic retrieval signals. For a query q and evidence item d,

$$s_{\text{hybrid}} = \alpha s_{\text{tfidf}} + (1 - \alpha) s_{\text{emb}},$$

where s_{tfidf} is TF–IDF cosine similarity and s_{emb} is embedding cosine similarity. The parameter $\alpha \in [0, 1]$ balances exact clinical terminology with semantic matching. The Cluster Profile is always included, and the remaining evidence is selected from the Snippet Bank according to the hybrid score, conditioning generation on a concise and relevant evidence set.

Evidence-constrained generation (Step 5). The language model receives the cluster profile and retrieved snippets and must justify statements using this evidence. Each major statement requires a citation; if support is lacking, the model outputs "insufficient evidence." Additional safeguards prohibit diagnostic or prescribing language. The summary follows a structured format including population overview, key indicators, and risk factors, producing an auditable description linked to observable patient data.

3. Experimental Results

This section evaluates whether the proposed cluster-aware RAG framework improves the reliability of medical summaries. First, we provide the *Experimental Protocol* paragraph in the Appendix A, which presents the experimental protocol, including the dataset, baselines, and evaluation methodology. Then, Section 3.1 reports and analyzes the results with respect to faithfulness, completeness, and safety, and Sec. 3.2 discusses the behavior of the model under evidence constraints.

3.1. Results, Evaluation, and Analysis

We evaluate cluster-level summarization under two conditions: true-label grouping (S1 → S1-RAG) and clustering-based grouping (S4 → S4-RAG). In all RAG experiments, summaries are generated using citation constraints and hybrid retrieval. The hybrid weight α is selected from $\{0, 0.25, 0.5, 0.75, 1\}$ based on the supported-claim rate, with ties broken by lower overreach. The best performance occurs at $\alpha = 0.25$, corresponding to 25% lexical retrieval and 75% semantic retrieval.

Faithfulness. Faithfulness is evaluated using the supported-claim rate, obtained by decomposing each summary into atomic claims and verifying whether each claim is justified by at least one retrieved evidence item (cluster profile statistics or report snippets). As shown in Table 1, retrieval grounding consistently improves factual reliability across all clusters. For the true-label grouping (S1), the macro-average supported-claim rate increases from 11.1% to 33.3% (+22.2 points), while for clustering-based grouping (S4) it increases from 8.3% to 33.3% (+25.0 points). The improvement

is most pronounced in the average- and high-risk clusters, where unsupported generalizations are frequent in baseline summaries.

Table 1. Faithfulness evaluation based on supported and unsupported claim rates for baseline and grounded summaries. Retrieval grounding consistently increases the proportion of evidence-supported statements across all clusters.

Comparison	Metric	Low	Average	High	Macro-Avg
S1 baseline	Supported-claim rate (%) ↑	33.3	0.0	0.0	11.1
S1-RAG	Supported-claim rate (%) ↑	41.7	33.3	25.0	33.3
Δ (S1-RAG − S1)	Supported-claim rate (%) ↑	**+8.3**	**+33.3**	**+25.0**	**+22.2**
S1 baseline	Unsupported-claim rate (%) ↓	66.7	100.0	100.0	88.9
S1-RAG	Unsupported-claim rate (%) ↓	58.3	66.7	75.0	66.7
Δ (S1-RAG − S1)	Unsupported-claim rate (%) ↓	**-8.3**	**-33.3**	**-25.0**	**-22.2**
S4 baseline	Supported-claim rate (%) ↑	25.0	0.0	0.0	8.3
S4-RAG	Supported-claim rate (%) ↑	41.7	33.3	25.0	33.3
Δ (S4-RAG − S4)	Supported-claim rate (%) ↑	**+16.7**	**+33.3**	**+25.0**	**+25.0**
S4 baseline	Unsupported-claim rate (%) ↓	75.0	100.0	100.0	91.7
S4-RAG	Unsupported-claim rate (%) ↓	58.3	66.7	75.0	66.7
Δ (S4-RAG − S4)	Unsupported-claim rate (%) ↓	**-16.7**	**-33.3**	**-25.0**	**-25.0**

The reduction in unsupported claims indicates a change in generation behavior rather than a stylistic difference. In the baseline setting, the model often produces plausible population-level statements inferred from general medical knowledge but not supported by the reports. After introducing retrieval grounding, statements must be justified by cluster statistics or representative patient narratives, shifting the summaries from inference-driven to evidence-driven descriptions. Notably, RAG summaries contain more extractable claims (approximately 12 per cluster versus 3–4 in the baseline) while still achieving higher support rates, suggesting that the improvement arises from better grounding rather than more conservative generation. Overall, cluster-aware grounding reduces hallucination-like behavior and yields summaries that are verifiable with respect to the underlying patient data.

Completeness. Completeness measures whether summaries explicitly mention the most clinically relevant findings, defined as the two most abnormal indicators in each cluster profile. As shown in Table 2, retrieval grounding preserves coverage in the low- and average-risk clusters and substantially improves it in the high-risk cluster. In both S1 and S4 settings, completeness for the high-risk group increases from 50% to 100%, and the macro-average coverage rises from 83.3% to 100%. This improvement reflects the role of structured evidence in guiding content selection. In the baseline setting, the model occasionally omits important abnormalities, particularly in clinically complex populations, because it relies on narrative patterns rather than explicit clinical indicators. Providing the cluster profile makes salient findings directly accessible during generation, encouraging the model to describe clinically significant conditions instead of generic observations. Consequently, grounding not only prevents unsupported statements but also reduces omission errors, producing summaries that are both verifiable and clinically informative.

Table 2. Completeness evaluation measured by coverage of clinically important indicators (must-mention findings). Grounded summaries achieve full coverage, particularly improving the high-risk cluster.

Comparison	Low	Average	High	Macro-Avg
S1 baseline	100.0	100.0	50.0	83.3
S1-RAG	100.0	100.0	100.0	100.0
Δ (S1-RAG − S1)	**+0.0**	**+0.0**	**+50.0**	**+16.7**
S4 baseline	100.0	100.0	50.0	83.3
S4-RAG	100.0	100.0	100.0	100.0
Δ (S4-RAG − S4)	**+0.0**	**+0.0**	**+50.0**	**+16.7**

Safety and overreach. Safety is evaluated using three rule-based indicators: diagnosis-like statements, medication recommendations, and absolute claims ("all", "always", "never"). As shown in Table 3, retrieval grounding eliminates clinically unsafe outputs. No diagnosis-like or prescribing statements appear in any RAG summary, and a medication recommendation observed in the high-risk baseline summary is removed after grounding. This indicates that citation-constrained prompting successfully prevents the model from producing clinical advice and keeps the summaries descriptive rather than prescriptive. However, RAG introduces a higher number of absolute claims

in the average- and high-risk clusters (5 and 3 flags, respectively), whereas the baselines contain none. This behavior suggests that grounding improves factual support but does not automatically control linguistic certainty: once evidence is available, the model tends to state supported observations more categorically. Therefore, grounding addresses factual correctness but not confidence calibration, and additional prompt constraints may be needed to discourage the use of universal quantifiers. The result highlights an important distinction between factual reliability and epistemic caution in clinical text generation.

Table 3. Safety analysis using diagnosis-like, medication, and absolute-claim flags. Grounded generation removes unsafe clinical advice but introduces more categorical phrasing in some clusters.

Comparison	Metric	Low	Average	High	Macro-Avg
S1 baseline	Absolute-claim flags (count) ↓	0	0	0	0.00
S1-RAG	Absolute-claim flags (count) ↓	0	5	3	2.67
Δ (S1-RAG − S1)	Absolute-claim flags (count) ↓	**0.0**	**+5.0**	**+3.0**	**+2.67**
S1 baseline	Medication flags (count) ↓	0	0	1	0.33
S1-RAG	Medication flags (count) ↓	0	0	0	0.00
Δ (S1-RAG − S1)	Medication flags (count) ↓	**0.0**	**0.0**	**-1.0**	**-0.33**
S1 baseline	Diagnosis-like flags (count) ↓	0	0	0	0.00
S1-RAG	Diagnosis-like flags (count) ↓	0	0	0	0.00
S4 baseline	Absolute-claim flags (count) ↓	0	0	0	0.00
S4-RAG	Absolute-claim flags (count) ↓	0	5	3	2.67
Δ (S4-RAG − S4)	Absolute-claim flags (count) ↓	**0.0**	**+5.0**	**+3.0**	**+2.67**
S4 baseline	Medication flags (count) ↓	0	0	1	0.33
S4-RAG	Medication flags (count) ↓	0	0	0	0.00
Δ (S4-RAG − S4)	Medication flags (count) ↓	**0.0**	**0.0**	**-1.0**	**-0.33**
S4 baseline	Diagnosis-like flags (count) ↓	0	0	0	0.00
S4-RAG	Diagnosis-like flags (count) ↓	0	0	0	0.00

3.2. **Discussion and Limitations**

Across the three evaluation dimensions, a consistent behavior emerges. Grounded generation improves factual support, increases coverage of clinically relevant findings, and removes unsafe outputs. The faithfulness results show that unsupported statements are substantially reduced when the model is constrained by retrieved evidence. At the same time, completeness improves, particularly in the high-risk cluster, indicating that the model not only avoids incorrect claims but also better captures salient abnormalities. The safety analysis further confirms that evidence-constrained prompting suppresses diagnostic and medication-related language, keeping the summaries descriptive rather than prescriptive. Without grounding, the model extrapolates population characteristics from internal medical knowledge, which can lead to plausible but unsupported generalizations and the omission of key indicators. With cluster-aware grounding, the model interprets explicit patient-derived evidence by using cluster statistics and representative reports to justify its descriptions. More broadly, the results indicate that reliable medical summarization requires traceability rather than only linguistic quality. In clinical use, summaries must be reviewable and defensible with respect to underlying patient data. The proposed framework reframes summarization as an evidence-grounded reporting task, producing outputs whose statements can be justified by observable evidence.

The present study has several limitations. First, evaluation is conducted on a relatively small synthetic cohort of 150 patients, which limits claims about real-world clinical generalization. Second, only one dataset is used, so the framework may still be partly adapted to the characteristics of this cohort. Third, the clustering component is fixed from prior work rather than compared against multiple clustering alternatives in this paper. These limitations mean that the current results should be interpreted as a controlled validation of the grounding mechanism rather than a definitive benchmark for clinical deployment.

4. **Conclusion**

This paper introduced a cluster-aware Retrieval-Augmented Generation framework for medical report summarization. The central idea is to generate summaries from explicit patient-derived evidence rather than from loosely structured reports alone. By organizing patients into clinically coherent populations and conditioning generation on both aggregated statistics and retrieved report

snippets, the proposed method produces summaries that are traceable to observable data. Experimental results on this controlled synthetic cohort indicate that grounding improves reliability under the studied evaluation setting. Across both true-label and clustering-based settings, supported-claim rates increase consistently while unsupported claims decrease, indicating reduced hallucination-like behavior. Grounding also improves completeness by ensuring that clinically important abnormalities are explicitly described, and safety constraints prevent diagnostic and medication-related language. Although grounded summaries exhibit lower lexical similarity to reference summaries, they provide verifiable and auditable descriptions, which are more appropriate for clinical applications. More broadly, the findings suggest that medical summarization should be treated as an evidence-grounded reporting task rather than a purely linguistic generation task. Future work will investigate evaluation on larger and real clinical datasets, comparison with additional clustering strategies, and improved calibration methods.

Acknowledgments

The authors acknowledge the use of ChatGPT (OpenAI) for assistance with language editing. The authors reviewed and edited all AI-generated content and take full responsibility for the final version of the paper.

References

[1] J. J. Liang, C.-H. Tsou, B. Dandala, A. Poddar, V. Joopudi, D. Mahajan, et al. "Reducing Physicians' Cognitive Load During Chart Review: A Problem-Oriented Summary of the Patient Electronic Record". In: *AMIA Annual Symposium Proceedings*. 2021, pp. 763–772.

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21 (2020), 140:1–140:67.

[3] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. "On Faithfulness and Factuality in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 1906–1919.

[4] E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, D. Pimenta, et al. "A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation". In: *npj Digital Medicine* 8 (2025), p. 274.

[5] K. Torabizadeh, R. Hedjam, M. Allaoui, and B. Abdulrazak. "Embedding-Enhanced Patient Clustering for Customized Medical Report Summarization using LLMs". In: *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*. 2025, pp. 1–6. DOI: 10.1109/ACDSA65407.2025.11165844.

[6] L. van der Maaten and G. Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[7] L. McInnes, J. Healy, N. Saul, and L. Großberger. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 10.21105/joss.00861.

[8] M. Allaoui, S. B. Belhaouari, R. Hedjam, K. Bouanane, and M. L. Kherfi. "t-SNE-PSO: Optimizing t-SNE using particle swarm optimization". In: *Expert Systems with Applications* 269 (2025), p. 126398.

[9] M. Allaoui, R. Hedjam, K. Bouanane, M. S. Allili, M. L. Kherfi, and S. B. Belhaouari. "Exploring Non-Negativity for Improved Manifold Embedding: Application to t-SNE". In: *Knowledge-Based Systems* (2025), p. 114547.

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. 2020.

[11] M. Alkhalaf, L. Alonazi, A. Alenazi, F. Almehmadi, D. P, A. Abdullateef, A. Abousalem, M. Ahmed, H. F. Alhakami, A. A. Alshumrani, and N. Saleh. "Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from EHR". In: *Journal of Biomedical Informatics* 156 (2024), p. 104662. DOI: 10.1016/j.jbi.2024.104662.

[12] S. Min, K. Krishna, X. Lyu, and et al. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023, pp. 12076–12100. DOI: 10.18653/v1/2023.emnlp-main.741.

[13] A. Nenkova and R. Passonneau. "Evaluating Content Selection in Summarization: The Pyramid Method". In: *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2004, pp. 145–152.

Appendix A. Experimental Protocol

We evaluate the proposed framework using the same synthetic cohort design as in our prior work due to the lack of publicly available clinical datasets suitable for controlled summarization analysis. The dataset contains approximately 150 patients stratified into low-, average-, and high-risk hypertension groups, each associated with structured clinical variables (e.g., blood pressure, BMI, glucose, and cholesterol) and a narrative clinical report. Because true population groups are known, the dataset allows separation of grouping errors from summarization errors in a controlled setting. The goal of this study is therefore to evaluate the effect of evidence grounding on summary reliability rather than to claim broad clinical generalization.

To isolate the effect of grounding, clustering is fixed to the best-performing configuration from our prior work [5] (t-SNE-PSO [8] followed by K-Means with $k = 3$), and clusters are mapped to their corresponding risk categories. K-Means is used here for two practical reasons: first, it provides stable and interpretable partitions after the embedding stage in our earlier pipeline; second, the cohort is organized around three clinically meaningful risk strata, making $k = 3$ a transparent choice for this controlled evaluation. We do not claim that K-Means is universally optimal; rather, it serves here as a simple and reproducible clustering component within the larger grounding framework. We compare the proposed Hybrid RAG method against non-grounded summarization under two settings: summaries generated from true-label groupings (S1 baseline) and from embedding-based clustering (S4 baseline), where S4 corresponds to the t-SNE-PSO cluster summaries introduced in our previous work [5]. Grounded variants (S1-RAG and S4-RAG) generate summaries using the cluster profile and retrieved evidence snippets.

In the present study, the same cohort is used to form clusters, construct the evidence artifacts, generate cluster-level summaries, and evaluate those summaries. Accordingly, the reported results should be interpreted as a within-cohort evaluation of grounding effectiveness rather than as out-of-sample performance on a separate held-out test set. Evaluation focuses on clinical reliability rather than textual similarity. Faithfulness is assessed by decomposing each summary into 8–15 atomic claims and measuring supported and unsupported claim rates using retrieved evidence [3, 12]. Completeness is measured as must-mention coverage of the two most abnormal indicators identified from the cluster profile [13]. Safety is evaluated by counting diagnosis-like statements, medication recommendations, and absolute claims (“all”, “always”, “never”) [4].