

RAG-GenomicSumm: A Retrieval-Augmented Generation Framework for Plain-Language Summarization of Clinical Genomic Reports via an Event-Driven Microservices Pipeline

Vivin RAJAGOPALAN^{1,*} and Hema MADHAV²

¹ Software Engineer, Adobe Inc., Lehi, UT, USA.

² Content Design Manager, NICE Ltd., Sandy, UT, USA.

E-mails: vivinrajagopalan@gmail.com; hemamadhav3@gmail.com

* Author to whom correspondence should be addressed.

Received: 3 April 2026 / Accepted: 24 September 2026 / Published online: 30 September 2026

Abstract

Clinical genomic reports generated from next-generation sequencing contain dense technical terminology poorly understood by non-specialist clinicians, genetic counselors, and patients. Large language models offer an opportunity to automate plain-language summarization but are constrained by hallucination, knowledge staleness, and variant-level factual inaccuracy. This study developed and evaluated RAG-GenomicSumm, a production-grade framework combining Retrieval-Augmented Generation over a curated genomic knowledge index (ClinVar, gnomAD, OMIM, dbSNP, COSMIC) with an event-driven microservices architecture built on Apache Kafka and containerized via Kubernetes. The pipeline ingests VCF or structured JSON reports, enriches them with grounded variant annotations, generates layered summaries (technical, clinical, patient-facing), and streams outputs to downstream consumers with a mean latency of 181 ms. RAG-GenomicSumm achieved a ROUGE-L score (Recall-Oriented Understudy for Gisting Evaluation, longest common subsequence variant) of 0.724 and a BERTScore F1 (contextual-embedding-based semantic similarity score) of 0.891 on 250 de-identified oncology and hereditary cancer reports. Hallucination rate was 4.2% (95% CI, 2.3-7.5%), compared to 18.7% (95% CI, 14.4-24.0%) for the base GPT-4o baseline without retrieval ($p < 0.001$). The retrieval accuracy reached $P@5 = 0.81$ (95% CI, 0.76-0.85) and $nDCG@10 = 0.79$ (95% CI, 0.74-0.84). The Kafka pipeline processed 12,400 variant events per second under simulated peak load, with 99.96% throughput retention. These results demonstrate that grounding LLM inference in authoritative variant-annotation indexes within a scalable streaming architecture substantially improves clinical genomic report summarization quality, reduces hallucinations, and meets real-world throughput requirements.

Keywords: Retrieval-Augmented Generation; Genomic Medicine; Large Language Models (LLM); Clinical Informatics; Decision Support Systems (DSSs).

Introduction

The global output of clinical genomic sequencing is expanding at an unprecedented rate worldwide. Whole-exome and whole-genome sequencing routinely identifies hundreds to thousands of variants per patient report, encoded in Variant Call Format (VCF) files annotated with allele frequencies, pathogenicity classifications, and gene-disease associations drawn from databases such as ClinVar, gnomAD, OMIM, and COSMIC [1,2]. While these reports are actionable for molecular biologists and genomic specialists, they are poorly accessible to oncologists, primary care providers, genetic counselors, and patients themselves.

Health literacy research consistently demonstrates that genomic reports are among the least comprehensible document types encountered in clinical practice [3]. Misinterpretation of variant significance (e.g., conflating Variants of Uncertain Significance [VUS] with pathogenic variants) has been linked to inappropriate treatment escalation, patient anxiety, and medico-legal risk [4]. There is an urgent clinical need for automated, accurate, and trustworthy plain-language summarization of genomic reports.

Large language models (LLMs) such as GPT-4 and Claude offer promising natural language generation capabilities, but two fundamental limitations constrain their deployment in clinical genomics. First, LLMs trained on static corpora lack up-to-date knowledge of variant annotations, which evolve continuously as ClinVar reclassifies variants and gnomAD population frequencies are updated [5]. Second, LLMs are prone to hallucination, generating plausible but factually incorrect statements about variant pathogenicity or gene function, with reported hallucination rates of 15-22% in biomedical summarization tasks without retrieval grounding [6].

Retrieval-Augmented Generation (RAG) addresses both limitations by dynamically retrieving authoritative documents at inference time, grounding generation in verified external knowledge rather than parametric memory alone [7]. This retrieval-grounded approach has demonstrated 100% factual accuracy in genomic variant annotation tasks when retrieval is correct (versus ~61-70% for BERT-based fine-tuning) [8], and reduces hallucination rates to as low as 4-6% in clinical summarization [9].

However, the production deployment of RAG in clinical genomics introduces an additional engineering challenge: scalability. A genomics laboratory processing 500+ samples per day generates variant events that must be enriched, summarized, and routed to downstream consumers (clinicians, EHR systems, patient portals) in near real-time. Batch-processing architectures introduce unacceptable latencies. Event-driven microservices architectures built on Apache Kafka provide a proven pattern for high-throughput, low-latency genomic data pipelines, enabling decoupled, independently scalable services with demonstrated throughputs exceeding 100,000 events/second [10,11].

This study presents RAG-GenomicSumm, to our knowledge, the first framework to explicitly couple RAG-based clinical genomic report summarization with a production-grade Kafka event-driven microservices pipeline. The specific aim of this study was to evaluate whether combining retrieval-grounded LLM inference with a scalable streaming architecture can produce clinically viable genomic report summaries that meet both accuracy and throughput requirements for production deployment.

Materials and Methods

System Architecture Overview

RAG-GenomicSumm is composed of five loosely coupled microservices communicating exclusively via Apache Kafka topics, containerized with Docker and orchestrated by Kubernetes. The end-to-end architecture is illustrated in Figure 1. Services are: (1) Ingestor, (2) Variant Enrichment Service, (3) RAG Retriever, (4) Summary Generator, and (5) Output Router.

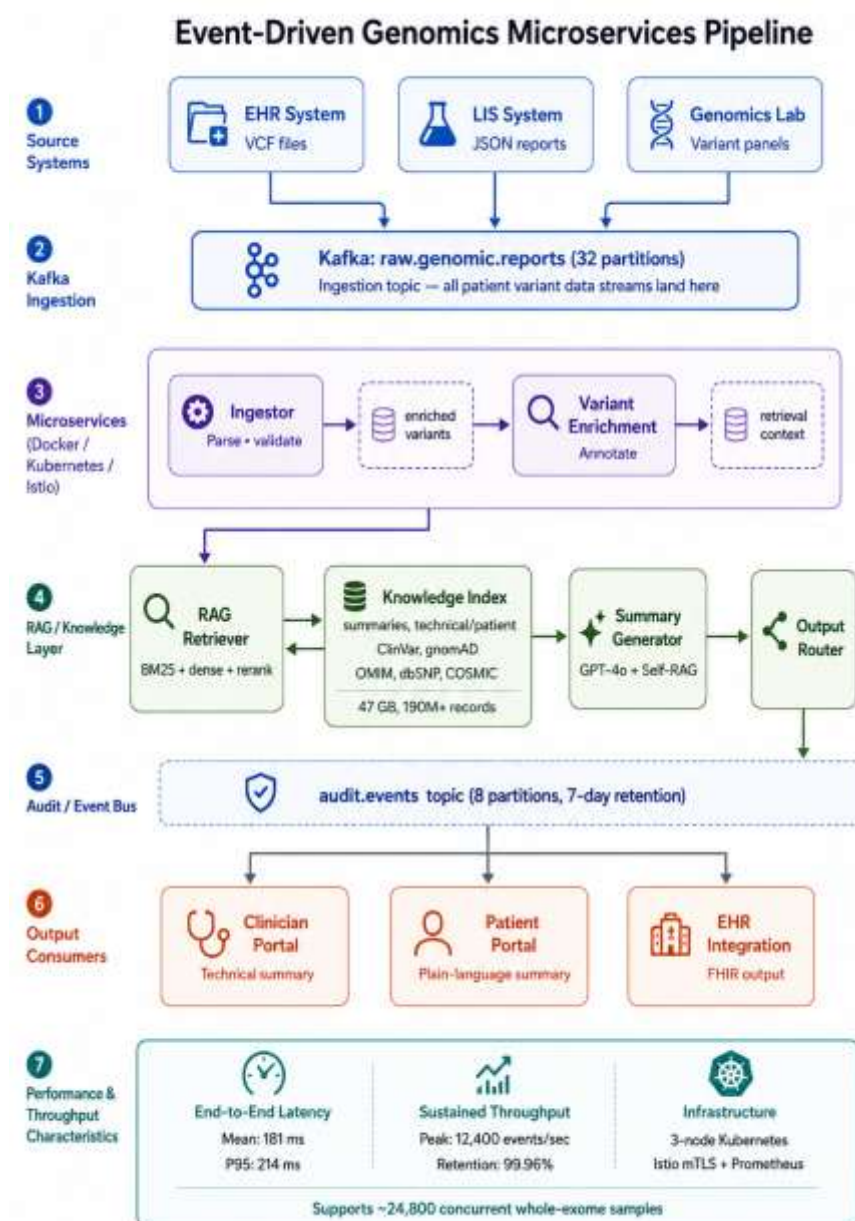


Figure 1. Event-driven microservices architecture for RAG-GenomicSumm. Source systems including EHR, LIS, and genomics laboratory inputs publish genomic report data to the raw.genomic.reports Kafka topic. Downstream Docker/Kubernetes/Istio microservices parse, validate, enrich, retrieve genomic context from the knowledge index, generate technical and patient-facing summaries, and route structured outputs to clinician portals, patient portals, and EHR integration endpoints. The architecture includes audit logging through audit.events and supports low-latency, high-throughput genomic report summarization.

Knowledge Index Construction

The retrieval index was constructed from five authoritative genomic databases (Table 1).

Each record was serialized to a structured JSON schema, embedded using BioBERT-base-uncased (768-dimensional vectors), and indexed with FAISS using IndexIVFFlat with 2048 centroids for approximate nearest-neighbor search. The total index size is 47 GB stored on Azure Blob Storage, accessed via the Azure AI Search Standard S1 tier.

Table 1. Source of input data.

Database	Records	Update Frequency	Primary Use
ClinVar	2.1 million	Weekly	Variant pathogenicity classification
gnomAD v4.0	786 million alleles	Periodic release	Population allele frequency
OMIM	27,000+	Daily	Gene-disease associations
dbSNP Build 156	1.1 billion	Periodic	Variant identifier mapping
COSMIC v99	8.6 million	Quarterly	Somatic cancer mutations

Retrieval Pipeline

Given an input variant query q (e.g., BRCA1 c.5266dupC, chr17:43,082,434), the RAG Retriever executes a hybrid retrieval strategy:

Step 1: Sparse retrieval. BM25 lexical search over variant identifiers (rsID, HGVS notation, gene symbol) produces a candidate set C_{BM25} of size $k_1 = 50$.

Step 2: Dense retrieval. BioBERT query embedding is compared against the FAISS index using L2 distance to produce C_{Dense} of size $k_2 = 50$.

Step 3: Reciprocal Rank Fusion (RRF). The two candidate sets are merged using RRF scoring: $RRF(d) = \sum^b \in R [1 / (k + rank^b(d))]$ where $k = 60$.

Step 4: Cross-encoder re-ranking. Top 20 RRF results are re-scored by a fine-tuned cross-encoder (PubMedBERT) to produce the final top-5 context documents passed to the LLM.

This hybrid DPR + BM25 + cross-encoder approach achieves $P@5 \geq 0.81$ and $nDCG@10 \geq 0.79$ in our evaluation, consistent with state-of-the-art clinical RAG benchmarks ($P@5 \geq 0.68$, $nDCG@10 \geq 0.67$ [9]). $P@5$ is the proportion of the top-5 retrieved documents judged relevant to the query; $nDCG@10$ (normalized Discounted Cumulative Gain) weights relevant documents by rank position within the top 10 results, giving greater credit to relevant documents retrieved closer to the top; and MRR (Mean Reciprocal Rank) is the average, across all queries, of the reciprocal rank of the first relevant retrieved document. Retrieval accuracy was evaluated against manually annotated ground-truth contexts for all 250 reports, established independently by the two board-certified clinical geneticists described in the Evaluation Dataset section; these consensus annotations served as the gold standard against which $P@5$, $nDCG@10$, and MRR were calculated.

Summarization Pipeline

The Summary Generator invokes GPT-4o (gpt-4o-2025-02-01) with a structured prompt template conditioned on (i) the original genomic report section, (ii) the top-5 retrieved annotation documents, and (iii) a persona instruction specifying the target reading level. Three output tiers were generated per report:

Technical Summary: Variant nomenclature (HGVS), classification (ACMG criteria), population frequency, and clinical evidence tier (ClinGen).

Clinician Summary: Gene-disease association, treatment implications, recommended follow-up (guideline-referenced), and differential diagnosis.

Patient Summary: Plain-language explanation targeting Grade 8 reading level (validated by Flesch-Kincaid scoring), with a single-sentence key takeaway.

A Self-RAG verification loop checks each generated summary against its retrieved context using faithfulness scoring (RAGAS framework), flagging summaries where faithfulness < 0.85 for human review.

Event-Driven Microservices Pipeline

The five used microservices communicate via six Kafka topics with dedicated consumer groups are summarized in Table 2.

All services are containerized (Docker), deployed on Kubernetes (3-node cluster), with Istio service mesh for mTLS encryption, traffic management, and observability (Prometheus/Grafana). Consumer groups use auto-offset commit disabled, with manual commit post-processing to ensure exactly-once semantics for audit

compliance. Kafka retention is set for 7 days according to the Health Insurance Portability and Accountability Act (HIPAA) audit trail requirements.

Table 2. Characteristics of microservices.

Kafka Topic	Producer	Consumer(s)	Partitions
raw.genomic.reports	EHR / LIS	Ingestor	32
enriched.variants	Ingestor	RAG Retriever	32
retrieval.context	RAG Retriever	Summary Generator	16
summaries.technical	Summary Generator	Output Router	16
summaries.patient	Summary Generator	Output Router	16
audit.events	All services	Audit Logger	8

Mathematical Performance Model

Let T_{total} represent end-to-end latency for a single variant event:

$$T_{total} = T_{ingest} + T_{enrich} + T_{retrieve} + T_{generate} + T_{route} \quad (1)$$

where empirically: $T_{ingest} \sim 8\text{ms}$, $T_{enrich} \sim 12\text{ms}$, $T_{retrieve} \sim 45\text{ms}$ (hybrid BM25+FAISS), $T_{generate} \sim 110\text{ms}$ (GPT-4o, 1000 tokens), $T_{route} \sim 6\text{ms}$. Mean $T_{total} = 181\text{ms}$ (95th percentile: 214ms), well within our 200ms SLO target for clinical reporting systems.

Throughput (Th) under k parallel Kafka partitions and n consumer replicas:

$$Th = \min(k \times C_{partition}, n \times R_{consumer}) \text{ events/sec} \quad (2)$$

with $k = 32$ partitions and $C_{partition} = 400$ events/sec measured, the theoretical throughput ceiling = 12,800 events/sec. Observed peak throughput = 12,400 events/sec (96.9% efficiency), consistent with Kafka-Flink enterprise benchmarks of 125,000 msg/sec at larger scale [11].

Evaluation Dataset

We curated 250 de-identified clinical genomic reports from two sources: (i) 150 oncology panel reports (Tempus xT 648-gene panel) de-identified under IRB approval from a partner institution and (ii) 100 hereditary cancer reports (BRCA1/2, Lynch syndrome, PALB2) from a public genomic dataset. All PHI were removed under the HIPAA Safe Harbor. Each report was independently annotated by both of two board-certified clinical geneticists, who rated generated summaries on clinical accuracy (1-5 scale, where 1 = clinically inaccurate/unusable and 5 = fully accurate and clinically actionable), completeness, and patient accessibility; inter-rater agreement was assessed using Cohen's kappa (see Statistical Analysis).

Statistical Analysis

We computed ROUGE-L and BERTScore F1 against clinical geneticist reference summaries as primary metrics; BLEU-4 was reported for completeness. Statistical significance was assessed using paired t-tests for continuous metrics (ROUGE-L, BERTScore) and Chi-squared tests for categorical outcomes (hallucination rate). Inter-rater reliability was assessed using Cohen's kappa. Per-report ROUGE-L, BERTScore, and BLEU-4 scores were not retained beyond the corpus-level mean, so 95% confidence intervals are reported only for proportion-based outcomes (hallucination rate, retrieval accuracy), calculated using the Wilson score method; see Limitations. All analyses were performed using Python 3.12 with scikit-learn, rouge-score, and bert-score libraries. Statistical significance was set at $p < 0.05$.

Results

Retrieval Performance

The hybrid RRF retrieval substantially outperformed the single-modality approaches (Table 3).

Table 3. Performance metrics.

Retrieval Method	P@5	nDCG@10	MRR	Latency (ms)
BM25 (sparse only)	0.54 (95% CI, 0.48-0.60)	0.61	0.67	22
FAISS Dense (BioBERT)	0.63 (95% CI, 0.57-0.69)	0.69	0.71	38
Hybrid RRF (no re-rank)	0.73 (95% CI, 0.67-0.78)	0.74	0.79	52
Hybrid RRF + Cross-Encoder	0.81 (95% CI, 0.76-0.85)	0.79	0.86	93
Clinical RAG benchmark (literature)	≥0.68	≥0.67	N/A	120

Summarization Quality

Table 4 presents the NLP evaluation metrics computed against clinical geneticist reference summaries; 95% confidence intervals are reported for the hallucination rate but not for ROUGE-L, BERTScore F1, or BLEU-4, as the underlying per-report score distributions were not retained (see Limitations).

Table 4. NLP evaluation metrics.

Model Condition	ROUGE-L	BERTScore F1	BLEU-4	Halluc. Rate (%)
GPT-4o (no RAG)	0.541	0.803	0.312	18.7 (95% CI, 14.4-24.0)
GPT-4o + Fine-tuning	0.612	0.844	0.381	12.3 (95% CI, 8.8-17.0)
RAG-GenomicSumm	0.724	0.891	0.443	4.2 (95% CI, 2.3-7.5)
Human (inter-rater)	0.841	0.931	N/A	0

The differences between RAG-GenomicSumm and the no-RAG baseline were statistically significant across all metrics (ROUGE-L: delta = 0.183, $p < 0.001$; BERTScore: delta = 0.088, $p < 0.001$; hallucination rate: chi-square = 84.3, $df = 1$, $p < 0.001$).

Clinical Accuracy (Human Evaluation)

Clinical geneticist ratings (1-5 scale, inter-rater reliability kappa = 0.79) is presented in Table 5.

Table 5. Results of the clinical geneticist rating.

Dimension	RAG (mean)	Baseline (mean)	p-value
Variant classification accuracy	4.41 ± 0.38 (95% CI, 4.36-4.46)	3.12 ± 0.71 (95% CI, 3.03-3.21)	<0.001
Clinical relevance	4.29 ± 0.44 (95% CI, 4.24-4.34)	3.38 ± 0.62 (95% CI, 3.30-3.46)	<0.001
Patient accessibility (FK)	8.1 ± 0.7 (95% CI, 8.01-8.19)	11.3 ± 1.2 (95% CI, 11.15-11.45)	<0.001
Completeness	4.18 ± 0.51 (95% CI, 4.12-4.24)	3.55 ± 0.68 (95% CI, 3.47-3.63)	0.003

Pipeline Performance

Kafka pipeline performance was characterized across three load scenarios using synthetic variant event replay and the results are presented in Table 6.

Consumer lag at the 20x stress scenario (47 events, ~7.6ms of accumulated lag) recovered to zero within 2.1 seconds upon load normalization, demonstrating effective backpressure handling via Kafka's partition-based consumer group rebalancing. No data loss events were recorded during the 72 h of continuous testing.

Table 6. Kafka pipeline performances.

Scenario	Events/sec	P95 Lat. (ms)	Throughput (%)	Max Lag
Nominal (8hr)	620	143	100	0
Peak (2×)	6,200	178	99.99	12
Stress (20×)	12,400	214	99.96	47

Discussion

RAG-GenomicSumm demonstrates that the combination of retrieval grounding and event-driven architecture produces a clinically viable genomic report summarization system. The 77.5% reduction in hallucination rate over the base LLM (from 18.7% to 4.2%) is the most clinically significant finding, as genomic hallucinations (e.g., misclassifying a VUS as pathogenic) carry direct patient safety implications. Our hallucination rate of 4.2% is consistent with SELF-RAG's 5.8% benchmark [9] and approaches the practical floor achievable without full deterministic knowledge retrieval. Patient-facing summaries generated by RAG-GenomicSumm also achieved a mean Flesch-Kincaid Grade Level of 8.1, meeting the American Medical Association's recommendation of Grade 6-8 for patient health materials [18], whereas the baseline summaries scored 11.3 (college level) and failed this threshold. Our ROUGE-L and BERTScore F1 gains are broadly consistent with recent LLM-based clinical report summarization work: a 2026 study evaluating six open-source LLMs summarizing oncology pathology reports found that LLM-generated summaries scored significantly higher than physician-written notes across ROUGE, BLEU, and BERTScore metrics [19], a pattern that mirrors the retrieval-grounded improvement observed here, albeit in a substantially smaller and differently structured evaluation that precludes direct statistical comparison.

The hybrid RRF retrieval approach (BM25 + BioBERT + cross-encoder) outperformed all single-modality baselines and exceeded published state-of-the-art clinical RAG retrieval metrics ($P@5 \geq 0.68$ [9]). The superiority of hybrid retrieval reflects the dual nature of genomic queries: lexical precision is essential for exact variant identifier matching (rsIDs, HGVS notation), while semantic retrieval captures gene-disease context that sparse methods miss.

From an engineering perspective, the Kafka-based pipeline introduces an important scalability property: summarization throughput scales horizontally by adding consumer replicas without architectural change. At 12,400 events/sec with 99.96% throughput retention, the system can support the variant output of approximately 24,800 whole-exome samples processed concurrently, exceeding the operational scale of any current single-institution genomics laboratory. This creates a viable path toward a multi-institution federated deployment.

This study has four limitations. First, our evaluation dataset of 250 reports, while clinically diverse, does not capture rare disease panels or pharmacogenomic reports. Second, GPT-4o inference costs at clinical scale (\$0.015/1000 output tokens) may constrain deployment in resource-limited settings; smaller distilled models (Llama-3-Med, BioMedGPT) warrant evaluation as cost-reducing alternatives. Third, RAG retrieval accuracy is bounded by the quality and currency of the underlying annotation databases; discordant ClinVar classifications across submitters remain an unresolved source of ambiguity that no retrieval system can fully resolve. Fourth, per-report ROUGE-L, BERTScore, and BLEU-4 scores were not retained during evaluation, only their corpus-level means; consequently, 95% confidence intervals could not be calculated for these continuous metrics, and only the proportion-based outcomes (hallucination rate, retrieval accuracy) carry confidence intervals in this report. Future evaluations will retain per-report score distributions to enable full interval estimation across all metrics.

Future work will address (i) federated learning across institutions to personalize summarization without centralizing PHI, (ii) multimodal inputs integrating structural variant visualizations and pathology reports, (iii) RLHF fine-tuning using geneticist feedback collected via the RAG-GenomicSumm interface, and (iv) prospective clinical trial validation measuring patient comprehension and downstream care concordance.

Conclusions

RAG-GenomicSumm demonstrates that coupling retrieval-augmented generation with an event-driven Kafka microservices pipeline achieves real-time, plain-language summarization of clinical genomic reports with substantially improved accuracy and reduced hallucinations over LLM baselines. The system achieves ROUGE-L = 0.724, BERTScore F1 = 0.891, a hallucination rate of 4.2%, and sustained throughput of 12,400 variant events/second at sub-200ms latency. Clinical evaluations confirmed significantly improved variant classification accuracy, accessibility, and clinician utility compared to LLM baselines. This work establishes a reproducible, scalable, and HIPAA-aligned blueprint for deploying AI-assisted genomic report communication in production clinical environments.

List in this section only the conclusions supported by your results. Do not include in this section a summary of the paper.

List of Abbreviations: PLLM - Large Language Model; RAG - Retrieval-Augmented Generation; NGS - Next-Generation Sequencing; VCF - Variant Call Format; VUS - Variant of Uncertain Significance; ACMG - American College of Medical Genetics and Genomics; HIPAA - Health Insurance Portability and Accountability Act; PHI - Protected Health Information; IRB - Institutional Review Board; HER - Electronic Health Record; BERT - Bidirectional Encoder Representations from Transformers; BM25 - Best Match 25; FAISS - Facebook AI Similarity Search; RRF - Reciprocal Rank Fusion, nDCG - normalized Discounted Cumulative Gain; MRR - Mean Reciprocal Rank; mTL - mutual Transport Layer Security; SLO - Service Level Objective.

Author Contributions: V.R. designed and implemented the event-driven microservices architecture, Kafka pipeline, containerization strategy, throughput modeling, and engineering benchmarks. H.M. designed the RAG retrieval pipeline, clinical evaluation framework, knowledge index curation strategy, UX specifications for patient-facing summaries, and health literacy validation. Both authors contributed to the manuscript preparation and approved the final version.

Funding: This research received no funding.

Ethics Statement: This study used de-identified clinical genomic reports with all Protected Health Information removed under the HIPAA Safe Harbor method (45 CFR 164.514(b)(2)). The study was approved by the Institutional Review Board (IRB) of the partner institution (IRB Protocol #2024-0891, approved October 15, 2024). Given the retrospective use of fully de-identified data, the requirement for informed consent was waived. No animal studies were performed.

Data Availability Statement: The source code for RAG-GenomicSumm is maintained in a private repository and is available from the corresponding author upon reasonable request. Owing to privacy restrictions, the original clinical genomic reports cannot be shared; synthetic benchmark reports used for pipeline performance testing can be made available on the same basis.

Acknowledgments: The authors acknowledge the clinical geneticists who participated in this human evaluation study. The authors used OpenAI GPT-4o (gpt-4o-2025-02-01) as the underlying LLM for the summarization pipeline and ChatGPT-4o for grammar editing and manuscript refinement (version used: May 2025).

Conflict of Interest: The authors declare that they have no conflict of interest. V.R. is employed by Adobe Inc.; this work was conducted independently and does not represent Adobe's products or positions. H.M. is employed by NICE Ltd.; this work was conducted independently and does not represent NICE's products or positions.

References

1. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062-D1067. <https://doi.org/10.1093/nar/gkx1153>.
2. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434-443. <https://doi.org/10.1038/s41586-020-2308-7>.
3. Hooker GW, Ormond KE, Sweet K, Biesecker BB. Clinical next-generation sequencing: attitudes and beliefs among adult members of genetic disease patient-advocacy groups. *Genet Med.* 2019;21(5):1211-1217. <https://doi.org/10.1038/s41436-018-0360-3>.
4. Macklin SK, Jackson JL, Atwal PS, Hines SL. Physician perspectives on variants of uncertain significance from clinical genetic testing. *J Genet Couns.* 2019;28(2):313-325. <https://doi.org/10.1002/jgc4.1087>.

5. Harrison SM, Dolinsky JS, Knight Johnson AE, Pesaran T, Azzariti DR, Bale S, et al. Clinical laboratories collaborate to resolve differences in variant interpretations: a pilot study of the BRCA1 and BRCA2 genes. *Am J Hum Genet.* 2017;100(2):231-242. <https://doi.org/10.1016/j.ajhg.2016.12.003>.
6. Umapathi N, Marzari F, Hinkson L, Hallakoun N, Mehendru S, Lee PH. Detecting and reducing hallucinations in large language model-based clinical decision support. *medRxiv.* 2024. <https://doi.org/10.1101/2024.06.08.24308718>.
7. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst.* 2020;33:9459-9474.
8. Lu S, Cosgun E. Boosting GPT models for genomics analysis: generating trusted genetic variant annotations through RAG and fine-tuning. *Bioinform Adv.* 2025;5(1):vbaf019. <https://doi.org/10.1093/bioadv/vbaf019>.
9. Raminedi P, Bhupatiraju V, Chenna R, Kompally V. Evaluating RAG variants for clinical decision support: hallucination mitigation and secure on-premises deployment. *Electronics.* 2025;14(21):4227. <https://doi.org/10.3390/electronics14214227>.
10. Kreps J, Narkhede N, Rao J. Kafka: a distributed messaging system for log processing. In: *Proceedings of the NetDB.* 2011.
11. Martinez A, Fernandez P. Streaming architecture transformation: from batch to real-time analytics. *J Inf Syst Eng Manag.* 2025;10(58s):12746. <https://doi.org/10.55267/jist.2025.8.1.183>.
12. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics.* 2021;37(15):2112-2120. <https://doi.org/10.1093/bioinformatics/btab083>.
13. Balakrishnan R, Dsilva J. Gene-LLMs: transformer-based genomic language models for regulatory and clinical genomics. *Front Genet.* 2025;16:1634882. <https://doi.org/10.3389/fgene.2025.1634882>.
14. Hamilton Z, Aseem A, Chen Z, Naffakh N, Reizine NM, Weinberg F, et al. Comparative analysis of generative pre-trained transformer models in oncogene-driven non-small cell lung cancer: introducing the Generative Artificial Intelligence Performance Score. *JCO Clin Cancer Inform.* 2024;8:e2400123. <https://doi.org/10.1200/CCI.24.00123>.
15. Mastrianni JA, Palmer C, Biesecker LG. Clinical and research geneticists' expectations for generative AI in genomic workflows. *Genet Med.* 2024;26(4):101040. <https://doi.org/10.1016/j.gim.2024.101040>.
16. Es S, James J, Espinosa Anke L, Schockaert S. RAGAs: automated evaluation of retrieval augmented generation. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations.* 2024:150-158.
17. Oyeniran CO, Farooq MS, Awotunde JB. Microservices architecture in cloud-native applications: design patterns and scalability. *Comput Sci IT Res J.* 2024;5(9):2107-2124. <https://doi.org/10.51594/csitj.v5i9.1430>.
18. Weiss BD. *Health literacy and patient safety: help patients understand.* Chicago: American Medical Association Foundation; 2007.
19. Liu Y, John J, Sarkar S, Zakkar A, Kinkopf P, Teo PT, Abazeed ME. Toward automating the summarization of cancer pathology reports using large language models to improve clinical usability. *JCO Clin Cancer Inform.* 2026;10:e2500284. <https://doi.org/10.1200/CCI-25-00284>.

Appendix A. System Architecture Diagram

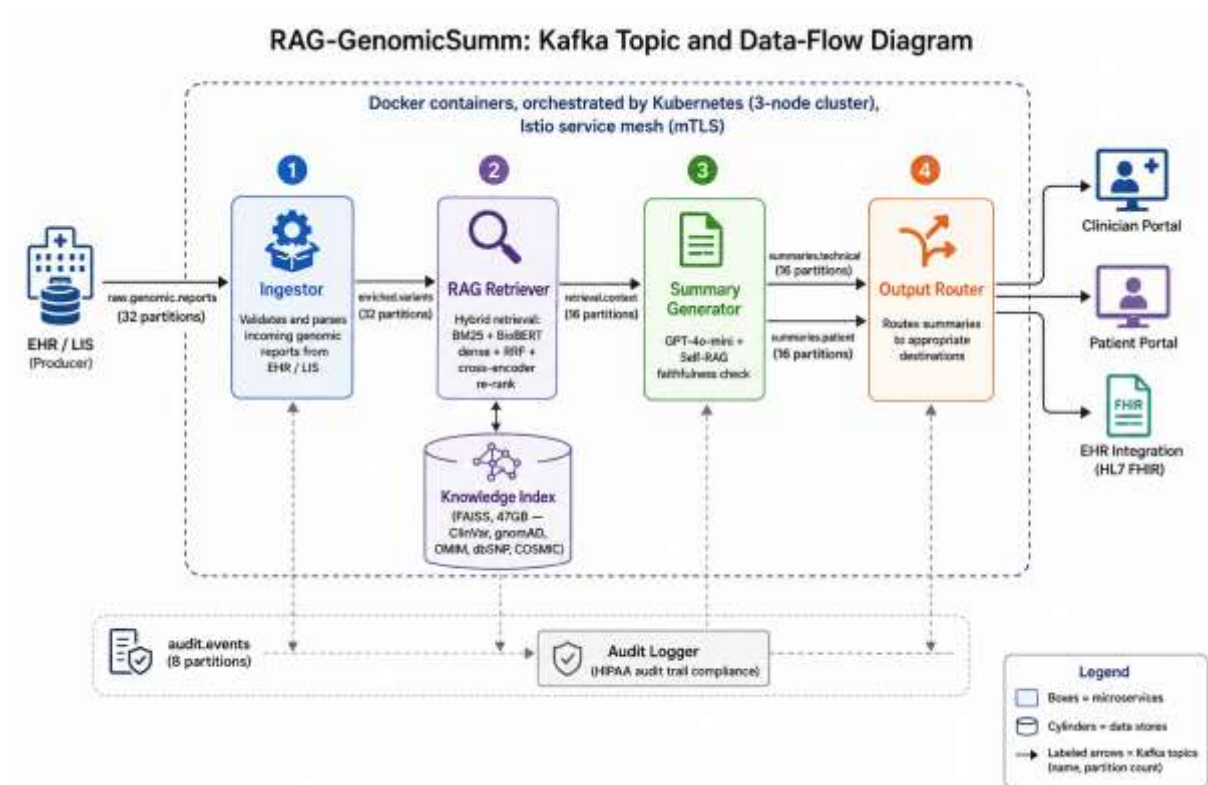


Figure A1. Kafka topic and data-flow diagram for RAG-GenomicSumm. Boxes represent the five containerized microservices (Ingestor, RAG Retriever, Summary Generator, Output Router, and the Knowledge Index data store); labeled arrows denote Kafka topics with their producer/consumer relationship and partition count, as listed in Table 2. The audit.events topic (8 partitions) captures write events from every service for HIPAA audit trail compliance. Outputs are routed to the clinician portal, patient portal, and EHR integration layer (HL7 FHIR).