A Deterministic Multi-Stage Retrieval Pipeline for Longitudinal EHR Question Answering

Shubham Agarwal[1], Thomas Searle[1], Ninoslav Majkic[2], Niko Muller-Grell[1], Richard Dobson[1,3]

[1]King's College London, London, UK
[2]South London and Maudsley NHS Foundation Trust, London, UK
[3]Health Data Research UK and University College London, London, UK
Correspondence: shubham.agarwal@kcl.ac.uk

Abstract

Retrieval-augmented generation (RAG) holds promise for clinical question answering over electronic health records (EHRs), but existing systems treat retrieval as an opaque subroutine, limiting auditability and reliability in patient care workflows. We introduce a deterministic multi-stage retrieval pipeline for longitudinal EHR question answering that decomposes retrieval into four distinct, ablated stages where each stage is instrumented with diagnostic metrics, making the flow of clinical evidence measurable and auditable at every step. Evaluated on a broad LLM-annotated cohort and an expert-annotated cardiovascular benchmark developed alongside clinicians from real ICU records, the full pipeline achieves 22-23% relative recall gain over a strong dense retrieval baseline across both cohorts, with consistent improvements in downstream answer quality. The pipeline's deterministic and transparent design addresses a critical gap in clinical NLP: retrieval systems that clinicians and researchers can not only rely on, but inspect, audit, and build upon for real-world deployment.

1 Introduction

Electronic Health Records (EHRs) capture rich longitudinal information about patients, including demographics, diagnoses, medications, and procedures, but a large fraction of this information exists only in unstructured free text: discharge summaries, nursing reports, radiology narratives, and echo reports (NHS, 2023; Häyrinen et al., 2008). Unlike coded fields, these narratives capture nuanced clinical reasoning, temporal relationships, and contextual detail that structured data systematically omits (Menachemi and Collum, 2011). A single patient admitted to an intensive care unit can accumulate hundreds of notes across multiple encounters, and over a longitudinal care history this figure reaches into the thousands (Zhang et al., 2017). The result is that clinically decisive information spanning areas such as trial eligibility (Raghavan et al., 2014), diagnosis (Shivade et al., 2014), and medication safety is present in the record but practically inaccessible at scale.

Conventional query interfaces and keyword search fail in this setting because they ignore clinical context, cannot resolve synonyms and abbreviations, and do not scale to long, redundant, and noisy documentation (Hill et al., 2021; Spasic et al., 2020). What clinicians and researchers need instead are systems that can answer natural language questions directly over a patient's record: whether a patient has ever met criteria for a particular syndrome (Pathak et al., 2013), how a clinical state has evolved across admissions (Silva and Matos, 2022), or why a medication was started, adjusted, or discontinued (Li et al., 2011). The core challenge is not data storage or extraction but reliable, traceable retrieval: identifying the specific note segments that support an answer, across a large and heterogeneous longitudinal record, without missing evidence that could affect patient care.

Retrieval-augmented generation (RAG) has emerged as a promising framework for this challenge, combining a retrieval component with a generative model conditioned on retrieved context (Lewis et al., 2020). In the EHR setting, RAG systems retrieve relevant passages from a patient's longitudinal notes and generate concise, grounded responses to clinical queries (Sivarajkumar et al., 2024; Alkhalaf et al., 2024), making the approach well suited to long, noisy, and heterogeneous documentation. However, the quality of generated answers depends directly on what is retrieved, as a generator conditioned on incomplete evidence will produce answers that are fluent but miss critical clinical information. Despite this, the question of how to engineer retrieval for high recall and auditability under realistic clinical constraints remains largely open.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 665–678

In this work, we address this gap with a deterministic multi-stage retrieval pipeline for longitudinal EHR question answering, designed around the constraints of real clinical workflows. The pipeline was developed and evaluated alongside clinical experts, with an expert-annotated cohort constructed specifically for cardiovascular medication review, a high-stakes setting where incomplete evidence recovery can directly compromise clinical decisions. Specifically, we:

- *Decompose retrieval into four distinct, ablated stages* that narrow the clinical corpus, expand query coverage, recover missed evidence, and optimise final ranking to ensure no clinically relevant information is lost.
- *Instrument each stage with diagnostic metrics*, ensuring the flow of clinical evidence is measurable and auditable at every step, enabling systematic improvement of the pipeline and, ultimately, reliable clinical use.
- *Validate on an expert-annotated cardiovascular cohort developed alongside clinicians from real ICU records and an LLM-annotated cohort*, achieving over 22% relative recall gain over a strong dense retrieval baseline across both cohorts with consistent improvements in downstream answer quality.

2 Related Work

Clinical question answering over EHRs requires systems to identify small sets of highly relevant, patient-specific evidence from large, heterogeneous, and longitudinal records. Unlike open-domain Question-Answering (QA), clinical settings impose strict requirements on evidence traceability, recall of critical findings, and bounded result sets suitable for clinician review since the retrieved evidence may directly affect patient care and must therefore be reliable and auditable (Roberts et al., 2017; Koopman et al., 2015). In this context, system reliability therefore depends not on answer fluency alone, but on whether the correct note segments are retrieved in the first place which is a challenge that has increasingly motivated retrieval-augmented approaches in clinical NLP.

Large Language Model (LLM) systems for clinical tasks frequently adopt RAG to ground outputs in external evidence (Lewis et al., 2020; Izacard and Grave, 2021; Guu et al., 2020). In medicine, such approaches have been applied to clinical question answering (Singhal et al., 2023; Lehman et al., 2023), summarization (Nazi and Peng, 2024), and decision support (Jin et al., 2021). However, evaluation in these systems typically emphasizes end-to-end answer correctness or textual quality, while the retrieval component itself receives little direct analysis.

Deterministic Information Retrieval (IR) approaches for EHRs often leverage medical ontologies such as SNOMED CT (Donnelly, 2006) and Unified Medical Language System (UMLS) (Bodenreider, 2004) to improve recall and interpretability (Sivarajkumar et al., 2024) . Named entity recognition and concept normalization systems such as cTAKES (Savova et al., 2010), MetaMap (Aronson, 2010), and MedCAT (Kraljevic et al., 2021a) have been widely used to attach structured semantics to clinical narratives. Ranking models, which score and order candidate passages by relevance to a query, range from traditional probabilistic methods such as BM25 (Robertson and Zaragoza, 2009) to neural dense retrievers adapted to biomedical text (Karpukhin et al., 2020; Lee et al., 2020; Gu et al., 2021). Despite these advances, most clinical IR pipelines remain relatively shallow, typically involving one or two retrieval stages followed by ranking, and sparse methods such as BM25 can struggle in clinical settings where the same concept can be expressed through varied terminology, abbreviations, and paraphrasing, motivating the shift towards dense retrieval methods and more flexible, LLM-driven retrieval frameworks.

Modern RAG systems increasingly embed retrieval inside LLM-driven control loops. In these architectures, LLMs may reformulate queries (Asai et al., 2024), generate intermediate hypotheses (Yao et al., 2023), judge evidence relevance (Asai et al., 2024), select passages (Yao et al., 2023), or determine when to stop retrieving (Jiang et al., 2023; Fan et al., 2024; Gao et al., 2023b). Variants of iterative (Shao et al., 2023; Trivedi et al., 2023), recursive (Kim et al., 2023) and agentic (Shinn et al., 2023) retrieval have been proposed to improve reasoning over multi-hop or ambiguous queries.

While LLM-driven retrieval introduces flexibility through model-guided query reformulation, passage selection, and iterative reasoning, it shifts control and evidence selection to opaque model-internal decisions, reducing auditability and in-

troducing variable latency and computational cost that complicate integration into time-sensitive and privacy-sensitive clinical workflows. Taken together, existing work presents two extremes: deterministic but relatively shallow clinical IR pipelines, and flexible but opaque LLM-orchestrated retrieval frameworks. What remains missing is a systematic treatment of retrieval itself as a multi-stage, diagnosable process for patient-level QA.

3 Methodology

3.1 Dataset

This study uses the MIMIC-III (Johnson et al., 2016) clinical database, a publicly available collection of deidentified electronic health records from intensive care units at the Beth Israel Deaconess Medical Center. MIMIC-III contains unstructured clinical documentation such as admission notes, progress notes, nursing notes, discharge summaries, radiology reports, and other narrative reports associated with ICU and related hospital encounters.

For this work, unstructured free-text notes from the NOTEEVENTS table were used, yielding longitudinal collections that reflect realistic documentation patterns across multiple admissions and care settings.

3.2 Problem Formulation

Each patient record consists of multiple hospital visits, and each visit comprises a collection of free-text clinical notes.
For a given patient i, the record can be written as:

$$R_i = \{V_{i1}, V_{i2}, \ldots, V_{iM_i}\},$$

where V_{im} denotes the m-th visit for patient i, and M_i is the number of visits. Each visit V_{im} is itself a set of notes

$$V_{im} = \{n_{im1}, n_{im2}, \ldots, n_{imK_{im}}\},$$

where n_{imk} is an unstructured free text clinical note associated with that visit, and K_{im} is the number of notes for visit m of patient i. Each note n_{imk} is associated with a clinical category c, where

$$c = \{\text{discharge, nursing, ECG, radiology, echo}\}.$$

The objective of this work is to enable patient-centered QA over these note collections. Given a natural language query q about a specific patient i, for a specific visit m, the system must produce a natural language answer a, and a set of supporting evidence segments E_i drawn from notes from one visit across one or multiple categories. In the RAG setting, a retriever g selects relevant note chunks and a generator h produces the final answer conditioned on them.

In our experiments, queries are evaluated under two complementary settings: a category-restricted setting, where E_i is limited to notes of a specific clinical category c, and a visit-level setting, where E_i spans all note types within a visit. Both settings reflect realistic clinical workflows and are evaluated throughout this work.

Notation note For clarity, we present the formalism for the category restricted setting within a single patient visit (i, m, c). The visit-level setting arises naturally by expanding the candidate corpus to $\bigcup_c \mathcal{C}_{imc}$. The proposed pipeline is not restricted to this scope, and cross-visit retrieval can be performed by expanding the candidate corpus to $\bigcup_m \mathcal{C}_{imc}$, without modifying the underlying retrieval or representation components.

3.3 Modules

3.3.1 Chunking

Clinical notes vary substantially in length and structure, and naïvely splitting them into fixed-size segments risks either breaking coherent clinical units or creating overly long passages. We adopt a structure-aware chunking strategy that respects natural note boundaries and category-level organisation.
For a given patient i and visit m, the visit-level note collection is $V_{im} = \{n_{im1}, \ldots, n_{imK_{im}}\}$, where each n_{imk} is an unstructured clinical note with an associated category c. Chunking operates at the level of a patient–visit–category triplet in all experimental settings as notes from different clinical categories are chunked independently since merging across categories would conflate clinically distinct documentation types and degrade retrieval signal. For a fixed (i, m, c), $\mathcal{N}_{imc} = \{n_{imc1}, \ldots, n_{imcK_{imc}}\}$ denote the sequence of notes of category c. From this sequence, we construct a sequence of chunks
$\mathcal{C}_{imc} = \{u_{imc1}, \ldots, u_{imcT_{imc}}\}$, where each chunk u_{imct} is a concatenation or sub-span of one or more elements of $\mathcal{N}_{imc}$ subject to a character-length constraint $L_{min} \leq |u_{imct}| \leq L_{max}$, with $L_{min} = 900$ and $L_{max} = 1500$.

This strategy, as outlined in Appendix A.1, outperforms purely length-based recursive segmen-

tation, as confirmed by ablations reported in Appendix A.2 (Table 3), and yields retrieval units that respect natural note boundaries, are easier to interpret clinically, and remain short enough to be handled efficiently.

3.3.2 Metadata Enhancement via NER

For each note chunk u_{imct}, we enrich the raw text with structured metadata derived from clinical Named Entity Recognition (NER). We apply the MedCAT (Kraljevic et al., 2021b) model to identify clinical entities (e.g., disorders, findings, medications) and leverage its built-in SNOMED CT (SNOMED) mapping to assign each detected entity e to an ontology concept with corresponding semantic category $s(e) \in \mathcal{S}$, where $\mathcal{S}$ denotes the space of clinical semantic types.

Formally, the NER model produces a set of entity annotations $\mathcal{E}_{imct} = \{(e_j, s(e_j))\}_{j=1}^{J_{imct}}$, where e_j is the entity span and $s(e_j)$ its semantic type. We then compute summary metadata for the chunk as

$$M_{imct} = \Big(\big(|\{e_j \in \mathcal{E}_{imct} : s(e_j) = \tau\}| \big)_{\tau \in \mathcal{S}}, \mathcal{E}_{imct} \Big)$$

which stores both entity counts per semantic type and the full list of detected entities. These metadata M_{imct} are stored alongside the chunk text u_{imct} in the retrieval index.

Attaching ontology-based categories to each chunk enables the pipeline to rapidly narrow a large clinical corpus to a semantically plausible subset before retrieval stages. For example, a query about medications can be directed towards chunks containing entities labeled as clinical drugs or medicinal substances, reducing noise and computational cost for downstream stages.

3.3.3 Embedding and Indexing

Each chunk, augmented with its MedCAT-derived metadata is encoded using a dense embedding model to produce a fixed-dimensional vector representation suitable for semantic similarity search (Karpukhin et al., 2020; Ni et al., 2022; Izacard and Grave, 2021). These embeddings are computed for all chunks in a patient's corpus and stored in a vector database that supports efficient approximate nearest-neighbor search. At query time, this index enables rapid retrieval of the most semantically similar chunks, serving as the foundation for the subsequent stages in the pipeline.

We use the `Chroma`[1] vector database and the `BAAI/llm-embedder` (Zhang et al., 2023) model as the encoder for both queries and chunks, as it has shown strong performance on semantic retrieval benchmarks.

[1] https://www.trychroma.com

3.3.4 Query Category Generation

To guide metadata filtering of the corpus, we generate relevant SNOMED CT semantic types from each query q. Using MedGemma 4B (Sellergren et al., 2025), a medically pre-trained language model, with few-shot prompting over $\mathcal{S}$, a set of candidate ontology categories curated from the SNOMED CT hierarchy in consultation with clinical experts, we produce two disjoint sets of semantic types: $\mathcal{S}_q^d \subseteq \mathcal{S}, \quad \mathcal{S}_q^c \subseteq \mathcal{S}$, where $\mathcal{S}_q^d$ denotes direct categories explicitly referenced in q (e.g., 'clinical drug', 'medicinal product' for medication queries), and $\mathcal{S}_q^c$ denotes complementary categories that frequently co-occur with direct types (e.g., 'dose form', 'frequency', 'unit of presentation'). The filtering instruction set enables nuanced corpus pruning by capturing both primary semantic types and their clinically associated aspects.

$$\mathcal{C}_{imc_q}^{\text{filter}} = \{u_{imct} \in \mathcal{C}_{imc_q} \mid \exists \tau \in \mathcal{S}_q : M_{imct}[\tau] > 0\}$$

This guides the metadata filter to select relevant notes, substantially reducing the retrieval corpus while preserving recall of relevant evidence.

3.3.5 Query Entity Extraction and HyDE

To refine filtering, we extract specific entities from query q using MedCAT NER restricted to categories $\mathcal{S}_q$:

$$\mathcal{E}_q^{\text{direct}} = \{e \in \mathcal{E}(q) \mid s(e) \in \mathcal{S}_q\}.$$

In parallel, Hypothetical Document Embeddings (HyDE) (Gao et al., 2023a) generates a hypothetical document d_q from q, that approximates what a relevant clinical note might look like, exposing semantic content and clinical terminology that may not be explicit in the original query phrasing. We use a MedGemma 4B model for this. On this document, we again run NER restricted to $\mathcal{S}_q$, yielding a set of candidate entities

$$\mathcal{E}_q^{\text{hyde, raw}} = \{e \in \mathcal{E}(d_q) \mid s(e) \in \mathcal{S}_q\}.$$

These candidates are then validated against entities observed in the notes corpus, retaining only those that exceed a frequency threshold determined empirically, giving a validated set $\mathcal{E}_q^{\text{hyde}} \subseteq \mathcal{E}_q^{\text{hyde, raw}}$.

The final query entity set is $\mathcal{E}_q = \mathcal{E}_q^{\text{direct}} \cup \mathcal{E}_q^{\text{hyde}}$, which we use to refine filtering:

$$\mathcal{C}_{imc_q}^{\text{filter+}} = \{u_{imct} \in \mathcal{C}_{imc_q}^{\text{filter}} \mid \mathcal{E}_q \cap \mathcal{E}_{imct} \neq \emptyset\}.$$

This entity-level refinement further narrows the filtered corpus, ensuring that subsequent dense retrieval operates over a highly focused and semantically coherent candidate set.

3.3.6 Semantic Retrieval

From the entity-refined filtered set $\mathcal{C}_{imc_q}^{\text{filter+}}$, we perform dense semantic retrieval to obtain an initial candidate set. All chunks across a patient's record are embedded using an encoder for all $u_{imct} \in \bigcup_{i,m,c} \mathcal{C}_{imc}$.

At query time, given query q with embedding $\mathbf{q}$, we compute cosine similarities over the refined filtered corpus and retrieve the top-k' chunks where

$$k' = \min\left(30, \max\left(1, \left\lfloor 0.3 \cdot |\mathcal{C}_{imc_q}^{\text{filter+}}| \right\rfloor\right)\right).$$

This adaptive k' scales with filtered corpus size (taking $\sim$30% of available chunks, capped at 30), ensuring proportional coverage while guaranteeing at least one candidate. These top-k' chunks form the primary evidence set that would typically feed directly into generation in standard RAG frameworks; we instead refine this further with classifier recovery and re-ranking.

3.3.7 Classifier-Based Recall Recovery

Dense retrieval alone may still miss clinically important evidence, especially in long, heterogeneous notes where relevant information can be phrased in diverse ways. To reduce such false negatives, we introduce a recall-oriented classifier that re-examines chunks initially treated as non-relevant.

Let $\mathcal{C}_{imc_q}^{+} \subseteq \mathcal{C}_{imc_q}^{\text{filter+}}$ denote the top-k' chunks from semantic retrieval, with excluded chunks

$$\mathcal{C}_{imc_q}^{\text{neg}} = \mathcal{C}_{imc_q}^{\text{filter+}} \setminus \mathcal{C}_{imc_q}^{+}$$

For the classifier $\phi(q, u_{imct})$, we use a BERT model (Devlin et al., 2019) (`bert-base-uncased`), which demonstrated stronger generalisation than clinical variants in preliminary experiments, to predict chunk-level relevance for $(q, u_{imct}) \in \mathcal{C}_{imc_q}^{\text{neg}}$. We revive chunks exceeding threshold γ (chosen to be 0.8):

$$\widetilde{\mathcal{C}}_{imc_q}^{+} = \mathcal{C}_{imc_q}^{+} \cup \{\, u_{imct} \in \mathcal{C}_{imc_q}^{\text{neg}} \mid \phi(q, u_{imct}) \geq \gamma \,\}$$

Training details, data statistics, and hyper parameters are reported in Appendix A.5.

3.3.8 Re-ranking

After semantic retrieval and classifier recovery, the expanded candidate set $\widetilde{\mathcal{C}}_{imc_q}^{+}$ contains both high-confidence chunks from dense retrieval and revived chunks from classification. To produce a compact, high-quality evidence set for generation, we apply final re-ranking using the `BGE-Reranker` model (Chen et al., 2024) over $\widetilde{\mathcal{C}}_{imc_q}^{+}$:

$$\mathcal{C}_{imc_q}^{\text{final}} =_k \left\{ r(q, u_{imct}) \;\middle|\; u_{imct} \in \widetilde{\mathcal{C}}_{imc_q}^{+} \right\},$$

where $r(q, u_{imct})$ denotes the BGE-Reranker score and k is the final retrieval budget. While semantic retrieval and classifier recovery deliberately cast a wide net to maximise recall, the reranker tightens the candidate set to the final retrieval budget k as described in Section 4.2.2, prioritizing the most query-relevant chunks for the generator.

4 Experiments

4.1 Setup and cohorts

This work evaluates retrieval performance on two complementary cohorts: an LLM-annotated cohort for broad pipeline evaluation, and an expert-annotated cardiovascular cohort constructed alongside clinicians. Both are constructed from de-identified EHR data from MIMIC-III.

4.1.1 LLM-annotated Cohort

This cohort is built by sampling patients at random from MIMIC-III to avoid introducing task-specific or phenotype-specific bias into the evaluation set. For each selected patient, MedGemma 27B generates query–answer-span triplets conditioned on the available record, yielding a patient-specific, visit-specific, and category-specific set of triplets. The spans are then mapped to gold evidence documents in the record, so that each query has both (i) an associated set of relevant documents and (ii) a target answer.

The dataset comprises 1300 query-document-answer triplets derived from 5 patients across 3 visits, with on average 25 documents per query of which 1-2 are gold evidence. For a given patient and visit, queries are generated separately for each note type c rather than over a combined document set. Rather than shallow coverage across many patients, this cohort is designed for depth, extensively interrogating each patient across all visit types and note categories to stress-test the pipeline against the full complexity of real longitudinal records.

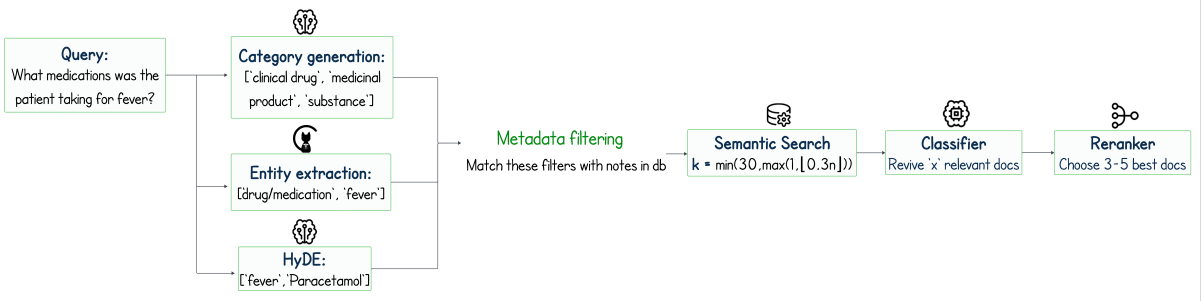

Figure 1: Overview of the four-stage retrieval pipeline

4.1.2 Expert-annotated Cardiovascular Cohort

The second cohort focuses on a clinically coherent subset of cardiovascular patients. Clinical experts first identify relevant patients and then construct 100 query–document mappings, covering typical information needs in cardiovascular medication review. For each patient–visit pair, experts identify relevant documents across all available note types. On average, there are 71 documents for each query, of which about 14 documents are annotated as gold evidence. This cohort therefore serves as the clinical validation benchmark, with high-quality, task-specific annotations against which the pipeline's retrieval performance can be assessed. Unlike existing EHR QA benchmarks that predominantly focus on factoid questions requiring a single supporting document, the average of 14 gold evidence documents per query here reflects the true complexity of real clinical information needs across heterogeneous note types. Appendix A.4 describes the annotation process. A calibration pilot on 30 queries achieved 95.8% inter-annotator agreement, confirming dataset reliability. Across the 100 queries, this dataset amounts to over 7400 document-level relevance judgments, representing a substantial expert annotation effort. This cohort directly reflects the cardiovascular medication review workflow described in Section 6.

Both the cohorts are evaluated under the setting natural to them: the *category-restricted* setting, where retrieval is limited to documents of a specific category c, is applied to the LLM-annotated cohort; and the *visit-level* setting, where retrieval spans all document types within a visit, is applied to the expert-annotated cohort.

4.2 Evaluation Metrics

Notation For a given query, C_{original} denotes all the document chunks in the patient record, C_{filtered} chunks that remain after filtering, G gold chunks associated with the query and C_{revived} the chunks reintroduced by the classifier.

4.2.1 Filtering stage metrics

At the filtering stage, we report:

Corpus ratio The fraction of chunks retained after filtering: $\frac{|C_{\text{filtered}}|}{|C_{\text{original}}|}$

Filtering recall The fraction of gold chunks that remain in the filtered corpus: $\frac{|G \cap C_{\text{filtered}}|}{|G|}$

Revived (classifier configurations only) The fraction of classifier revived chunks that are true gold evidence, reported as 'A/B': $|G \cap C_{\text{revived}}|/|C_{\text{revived}}|$.

4.2.2 Retrieval stage metrics

At the semantic search or reranking stage, operating on the filtered corpus and returning k chunks per query, we report:

Recall@k The fraction of gold chunks appearing among the top-k retrieved chunks.

Averaging All metrics are macro-averaged over queries, so each query contributes equally regardless of the number of associated gold chunks.

Choice of k. We set k=3 for the LLM-annotated cohort (around 25 candidate documents, 1-2 gold) to reflect a low-budget retrieval scenario, and k=20 for the expert-annotated cohort (around 71 candidate documents, 14 gold) to ensure adequate coverage.

4.2.3 Generation stage metrics

We report **ROUGE-L** (Lin, 2004), **BERTScore** (Zhang et al., 2019), and **SARI** (Xu et al., 2016) as generation stage metrics for the LLM-annotated cohort only, as the expert-annotated cohort does not provide reference answers for all queries.

5 Experimental Results

5.1 Baseline Retriever

Our primary baseline is a dense semantic search system that retrieves the top k document chunks per query from the patient corpus, without any additional stages. We use a fixed encoder to obtain embeddings for both queries and document chunks and rank chunks by cosine similarity. This standard RAG retriever configuration serves as the reference system for all subsequent comparisons, representing the most widely deployed retrieval approach in clinical RAG systems.

Baseline configuration The baseline retriever uses the configuration that achieved the best performance across preliminary ablations on chunk strategy, chunk size, and embedding model with full details of the compared methods reported for the LLM-annotated cohort in Appendix A.2.

5.2 LLM-annotated cohort

Table 1 reports retrieval results on the LLM-annotated patient cohort, showing the contribution of each individual component as discussed in Section 5.2.1. The full pipeline shrinks the search space to 78% of original size while boosting recall@3 from 0.60 to 0.73.

5.2.1 Pipeline Ablations

We ablate the pipeline cumulatively, adding metadata filtering, HyDE query expansion, a classifier, and a reranker in sequence, reporting filtering-stage metrics (corpus ratio, filtering recall, revived gold chunks) and recall@3 at each stage.

Metadata filtering Metadata filtering removes more than one quarter of the corpus (0.73), which is attractive from an efficiency standpoint, but the resulting filtering recall of 0.84 is still below the level needed for a reliable end-to-end system. Part of the difficulty is that metadata signals do not fully capture where fine-grained clinical facts are documented, and relevance is highly query dependent, so notes that appear generic or low-priority can still contain critical gold evidence and are occasionally pruned. This motivates the addition of beyond query-aware components.

HyDE HyDE query expansion meaningfully improves both filtering recall and recall@3 when added to the metadata filtered pipeline, raising filtering recall from 0.84 to 0.88 and recall@3 from 0.62 to 0.65. This confirms that queries enriched with clinically plausible terminology surface evidence that the original query phrasing alone misses. However, a filtering recall of 0.88 indicates a non-trivial fraction of gold evidence is missed, motivating a subsequent classifier to recover these.

Classifier The classifier aimed for recall recovery meaningfully boosts filtering recall from 0.88 to 0.94 and recall@3 from 0.65 to 0.69, while reviving 134 gold chunks previously excluded. This demonstrates that learned relevance assessment can correct the over-pruning of borderline cases for generic-looking notes that contain critical clinical facts for which metadata signals alone are insufficient. Together, the three stages build on each other, resulting in substantially stronger evidence recovery.

Reranker The reranker boosts recall@3 from 0.69 to 0.73 by reordering the candidate set, revealing that residual errors reflect fine-grained relevance distinctions among superficially relevant chunks rather than fundamental misses. This confirms the pipeline's strength: each stage corrects a distinct failure mode, culminating in a compact, high-precision evidence set for the generator.

5.3 Expert-annotated Cardiovascular Cohort

The expert-annotated cohort is substantially harder: 71 documents per query versus 25, and 14 gold evidence documents versus 1-2, spanning heterogeneous note types. This is not a controlled retrieval task but a genuine clinical information need, where evidence is distributed and missing any part of it has direct implications for medication review decisions. Despite this complexity, the pipeline delivers a 23% relative recall gain over the dense retrieval baseline, closely matching the +22% observed on the LLM-annotated cohort. The consistency of this gain across two fundamentally different evaluation settings, one broad and automatically annotated and one clinically curated and expert-validated, confirms that the pipeline's improvements are not an artefact of the evaluation conditions but reflect genuine, generalisable evidence recovery. These results demonstrate the pipeline's promise as a practical retrieval tool for real-world clinical use.

5.4 Generation Stage Evaluation

While our primary contribution is retrieval engineering, we include generation metrics to confirm retrieved evidence translates to improved end-to-end QA (Table 2). Using MedGemma 4B

Note: Each approach builds cumulatively on all prior stages

Table 1: Ablated retrieval results under natural configurations

Approach	Corpus Ratio	Filtering Recall	Revived	Recall@k
LLM-annotated cohort (category-restricted)				
Baseline	1.00	–	–	0.60
Metadata filtering	**0.73**	0.84	–	0.62
HyDE	0.75	0.88	–	0.65
Classifier	0.78	**0.94**	134/514	0.69
Reranker (Full Pipeline)	0.78	**0.94**	**376/1369**	**0.73**
Expert-annotated cohort (visit-level)				
Baseline	1.00	–	–	0.47
Metadata filtering	**0.65**	0.77	–	0.52
HyDE	0.66	0.80	–	0.52
Classifier	0.68	**0.83**	21/71	0.56
Reranker (Full Pipeline)	0.68	**0.83**	**38/91**	**0.58**

Table 2: Generation stage metrics

Metric	Baseline	Full pipeline
ROUGE-L	0.362	0.412
BERTScore	0.469	0.533
SARI	74.92	82.45

conditioned on top-k retrieved chunks, the full pipeline improves all generation metrics over baseline retrieval: ROUGE-L: +14% relative gain, BERTScore: +14% relative gain and SARI: +10% relative gain. These consistent gains across lexical overlap (ROUGE-L), semantic similarity (BERTScore), and system output quality (SARI) validate that improvements propagate to answer quality. Our focus on deterministic retrieval provides the strong evidence foundation on which generation quality directly depends.

6 Clinical Workflows

The pipeline is designed for real-world clinical deployment, with two immediate target workflows:

- *For clinical risk scoring*, such as CHA_2DS_2-VASc score (Lip et al., 2010) computation, where each variable must be evidenced by specific documentation across heterogeneous note types, high recall is the critical requirement, as missing a single signal can produce an incorrect risk score. The pipeline's multi-stage design ensures that evidence scattered across note types is systematically recovered, reducing the risk of incorrect risk scoring due to missed clinical signals.
- *For cardiovascular medication review*, clinicians must synthesize evidence distributed across discharge summaries, nursing notes, ECG reports, and echo reports. The pipeline surfaces this evidence in response to natural language queries, enabling clinicians to interrogate a patient's longitudinal record directly. The pipeline's deterministic, stage-wise design ensures every retrieval decision remains traceable and auditable, properties as important as performance in settings where the provenance of evidence directly affects clinical accountability.

7 Conclusion

Reliable clinical question answering over EHRs depends fundamentally on retrieval quality, yet the retrieval component has remained the least scrutinized part of clinical RAG systems. This work addresses that gap directly. By decomposing retrieval into four distinct, instrumented stages, the pipeline makes evidence recovery measurable and auditable at every step, achieving 22-23% relative recall gain over a strong dense retrieval baseline across cohorts. The consistency of this gain across two evaluation settings of fundamentally different complexity suggests genuine, generalisable improvements in evidence recovery rather than artefacts of evaluation design. Our work delivers a retrieval system that clinicians and researchers can rely on, inspect, understand, and build upon for real-world clinical use.

Limitations and Future Work

The LLM-annotated cohort was designed to prioritise depth over breadth, extensively interrogating five patients across all visit types, note categories, and query distributions to stress-test the pipeline against the full complexity of real longitudinal records. While this design choice ensures thorough within-patient evaluation, future work should extend coverage to a broader patient population to assess robustness across variation in documentation style and clinical complexity.

Similarly, evaluation is currently restricted to MIMIC-III, a single US medical centre; applying the pipeline to other EHR systems, clinical domains, languages, and documentation cultures is a natural and important next step.

Generation metrics are reported for MedGemma 4B only, and evaluating downstream answer quality under a broader range of generators is a straightforward extension. The pipeline's recall ceiling is tied to MedCAT's NER performance, and future work could explore more clinical NER systems or LLM-based entity extraction.

The classifier is trained on a heavily imbalanced dataset reflecting realistic clinical sparsity, and evaluating threshold robustness across note types and query distributions beyond those seen in training remains an open question. Finally, the multi-stage design introduces additional latency over a standard dense retrieval baseline; detailed runtime metrics and latency characterization for time-sensitive clinical deployment will be reported in future work.

Acknowledgments

This work was supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. SA, TS, RD are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RD is also supported by The National Institute for Health Research University College London Hospitals Biomedical Research Centre. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics

This work uses de-identified patient data from MIMIC-III under standard data use agreements. All clinical annotations were performed by qualified experts following structured guidelines. The pipeline is intended to support clinical decision-making, not replace clinical expertise, and retrieval outputs should always be reviewed by a clinician before informing patient care. The deterministic, auditable design of the pipeline is a deliberate ethical choice, ensuring that retrieval decisions remain transparent and traceable rather than opaque. We acknowledge the risk of retrieval bias, and note that evaluation on a broader and more diverse patient cohort remains important future work.

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.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alan Aronson. 2010. An overview of metamap: Historical perspective and recent advances. *JAMIA*, 17(3):229–236.

Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.

Olivier Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Kevin Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279–290.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Yu Gu, Robert Tinn, Hao Cheng, and et al. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*.

Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Jordan R Hill, Shyam Visweswaran, Xia Ning, and Titus K Schleyer. 2021. Use, impact, weaknesses, and advanced features of search functions for clinical use in electronic health records: a scoping review. *Applied Clinical Informatics*, 12(03):417–428.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, and et al. 2023. Active retrieval augmented generation. In *EMNLP*.

Di Jin, Eileen Pan, Nathalie Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, and et al. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009.

Bevan Koopman, Guido Zuccon, Anthony Nguyen, David Vickers, and Liaqat Butt. 2015. Information retrieval as a tool for clinical decision support: A systematic review. *JAMIA*, 22(4):799–811.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, and et al. 2021a. Medcat: The medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021b. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, and et al. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. 2023. Do we still need clinical language models? *EMNLP*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Ying Li, Hojjat Salmasian, Rave Harpaz, Herbert Chase, and Carol Friedman. 2011. Determining the reasons for medication prescriptions in the ehr using knowledge and natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2011, page 768.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Gregory YH Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry JGM Crijns. 2010. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2):263–272.

Nir Menachemi and Taleah H Collum. 2011. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, pages 47–55.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

NHS. 2023. Purpose of the gp electronic health record. Accessed on March 18, 2024.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and 1 others. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Jyotishman Pathak, Abel N Kho, and Joshua C Denny. 2013. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, 20(e2):e206–e211.

Preethi Raghavan, James L Chen, Eric Fosler-Lussier, and Albert M Lai. 2014. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Summits on Translational Science Proceedings*, 2014:218.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the trec 2017 precision medicine track. In *The... text REtrieval conference: TREC. Text REtrieval Conference*, volume 26, pages https–trec.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Guergana Savova, James Masanz, Philip Ogren, and et al. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes). *JAMIA*, 17(5):507–513.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, and et al. 2023. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230.

João Figueira Silva and Sérgio Matos. 2022. Modelling patient trajectories using multimodal information. *Journal of biomedical informatics*, 134:104195.

Karan Singhal, Shekoofeh Azizi, Tao Tu, and et al. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical information retrieval: a literature review. *Journal of healthcare informatics research*, 8(2):313–352.

SNOMED. Snomed international. Accessed on March 27, 2024.

Irena Spasic, Goran Nenadic, and 1 others. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval

with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Shunyu Yao, Jeffrey Zhao, Dian Yu, and et al. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

Rui Zhang, Serguei VS Pakhomov, Elliot G Arsoniadis, Janet T Lee, Yan Wang, and Genevieve B Melton. 2017. Detecting clinically relevant new information in clinical notes across specialties and settings. *BMC medical informatics and decision making*, 17(Suppl 2):68.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 Algorithm for Chunking

Algorithm 1 Chunking clinical notes

Require: Notes grouped by patient, admission, and category
Require: Minimum chunk size L_{min}, maximum chunk size L_{max}
Ensure: A list of text chunks with timestamps

```
Initialize empty list of chunks C
for each patient–admission–category group G do
    Sort notes in G chronologically
    Start an empty chunk buffer B, length L ← 0
    for each note in order do
        Let ℓ be the length of this note
        if ℓ is very large then
            Split the note into smaller pieces using recursive splitter
            for each piece do
                if adding it to B would exceed L_max then
                    Save current chunk to C and start a new one
                end if
                Add piece to current chunk and record its timestamp
            end for
        else if adding the note would exceed L_max then
            Save current chunk to C
            Start a new chunk with this note
        else
            Add the note to the current chunk
        end if
    end for
    if the final chunk buffer is not empty then
        Save it to C
    end if
end for
return all chunks C
```

A.2 Baseline Design Choices

Before introducing the full multi-stage pipeline, we first investigate two design choices for the baseline retriever: chunk size and embedding model. These experiments allow us to select a strong and well-justified baseline configuration, which we then use consistently in the main results and pipeline ablations.

A.2.1 Chunking Strategy and Size

We conduct an ablation study on different chunking strategies and chunk sizes to determine the most effective configuration for the baseline retriever. The chunking strategies compared are:

- **Recursive chunking:** Fixed-length recursive segmentation with partial overlap between segments.
- **Smart structural chunking (proposed):** A structure-aware approach that respects document boundaries and clinical section headers while maintaining approximately uniform chunk length.

For each strategy, we experimented with multiple target chunk sizes (approximately 400, 1200, and 1800 characters) and evaluated recall@3 using the dense retriever. The results, summarised in Table 3, show that the smart structural chunking strategy consistently outperforms purely length-based segmentation. A chunk size of around *1000* offered the best trade-off between retrieval recall and corpus compactness and was selected for all subsequent experiments.

We note that more sophisticated approaches such as semantic chunking or sentence-level segmentation were not explored in this work. In the clinical domain, such methods often disrupt the logical flow of medical narratives, where the order of statements, temporal markers, and contextual dependency carry significant meaning. To preserve the semantic continuity and clinical validity of the text, we retain structure-aware, non-semantic chunking methods.

A.3 Embedding Model

Using the optimal chunking configuration identified above, we compared several embedding models for the baseline retriever, including `nomic-ai/nomic-embed-text-v1.5` (Nussbaum et al., 2024), `intfloat/e5-base-v2` (Wang et al., 2022), `mixedbread-ai/mxbai-embed-large-v1` (Lee et al., 2024), `emilyalsentzer/Bio_ClinicalBERT` (Alsentzer et al., 2019) and `sentence-transformers/all-MiniLM-L6-v2` (Thakur et al., 2021). Each model was used to encode both query and chunk embeddings,

and performance was measured using recall@3. The results, shown in Table 4, demonstrate that `llm-embedder` provides the highest recall while maintaining computational efficiency. It was therefore selected as the default embedding model for both the baseline and all full-pipeline experiments.

Strategy	Chunk size	Recall@3
Recursive chunking	400	0.41
	1200	0.53
	1800	0.56
Smart chunking	(200, 500)	**0.45**
	(900, 1500)	**0.60**
	(1700, 2200)	**0.59**

Table 3: Chunk size ablation

Model	Recall@3
llm-embedder	**0.6**
nomic-embed-text-v1.5	0.59
e5-base-v2	0.58
mxbai-embed-large-v1	0.59
Bio_ClinicalBERT	0.42
all-MiniLM-L6-v2	0.55

Table 4: Embedding model ablation

A.4 Expert Annotation Process

Two clinical experts performed annotation for the cardiovascular cohort using a structured process:

Annotation guidelines We created a comprehensive annotation document detailing:

- **Task definition**: Identify all documents containing evidence for clinical information needs
- **Relevance criteria**: Direct mentions, inferable evidence, multi-hop reasoning requirements
- **Annotation procedure**: Consensus resolution, temporal judgment guidelines

Annotation tool We built a custom web interface to streamline the process, featuring:

- Side-by-side patient record browser (visit-organized notes by category)
- Query-document relevance marking
- Export to structured gold format for evaluation

Pilot and calibration An initial pilot on the same 30 queries achieved 95.8% inter-annotator agreement. A calibration meeting refined edge cases:

- Reasoning "hops": single-hop (direct) vs. multi-hop (synthesis across mentions)
- Temporal ambiguity in nursing notes (past vs. present conditions)
- Cross-document evidence distribution across note types

Remaining 70 queries were split among the annotators.

A.5 Classifier Training Details

Training data consists of query-chunk pairs generated using Gemma 12B (Team et al., 2025) over a held-out set of MIMIC-III patients with no overlap with either evaluation cohort. Chunks containing gold evidence serve as positive examples and non-relevant chunks within the same patient-visit-category serve as negatives, yielding 16894 training pairs in total with 2238 positive and 14656 negative, reflecting the realistic sparsity of relevant evidence in clinical notes. The classifier is trained with binary cross-entropy loss with class weights for 10 epochs, with learning rate 5×10^{-5} and weight decay 0.01.