Supplementary Material for RAGtio: A Modular Framework for Systematic Evaluation of Hybrid Retrieval-Augmented Generation Pipelines

Annamaria Defilippo[1], Nicola Procopio[2], Pietro Hiram Guzzi[1,*], Pierangelo Veltri[3], Patrizia Vizza[3]

[1]Department of Surgical and Medical Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy

[2]Technology Innovation Hub, AC Software, Lamezia Terme, Italy

[3]DIMES, University of Calabria, Rende, Italy

{annamaria.defilippo,hguzzi}@unicz.it

nicola.procopio@acsoftware.it

{patrizia.vizza,pierangelo.veltri}@dimes.unical.it

[*]Corresponding author: Pietro Hiram Guzzi; email: hguzzi@unicz.it

Supplementary Material

This file contains the two appendices removed from the main manuscript and reorganised as supplementary sections. The original source boundaries are retained through inline-file markers for editorial traceability.

S1 System Architecture Details

S1.1 Service Dependencies

Service	Technology	Role
`qdrant`	Qdrant v1.13.6	Persistent vector store (dense + sparse indices)
`ollama`	Ollama latest	Local LLM inference server
`app`	FastAPI + Uvicorn	REST API layer, pipeline orchestration
`ollama-init`	one-shot script	Bootstrap: pulls the target model at startup

Table S1: Docker Compose services and their roles in the RAGtio deployment.

All services communicate on a private bridge network (`rag_network`). The application container mounts `config.yaml` read-only and exposes evaluation reports through a named Docker volume.

S2 Indexing Pipeline

Source: `app/indexing.py` **Haystack primitive:** `Pipeline` (synchronous)
Purpose: Transform raw documents into embedded, semantically indexed chunks stored in Qdrant.

S2.1 Chunking Strategies

Chunking is the most critical design decision in a RAG system. It determines retrieval granularity and directly impacts both precision and recall. Three strategies are available, selectable via `indexing.chunking.strategy`.

S2.1.1 Character-level Split (`strategy: "character"`)

Splits text at a fixed character count with an optional overlap window:

$$\underbrace{c_1 \dots c_s}_{\texttt{chunk\_size}} \quad \underbrace{c_{s-o} \dots c_s}_{\texttt{overlap}} \underbrace{c_{s-o+1} \dots c_{2s-o}}_{\texttt{chunk\_size}} \quad \dots$$

Implementation: `RecursiveDocumentSplitter(separators=[""])`

Parameters:

- `chunk_size` (default: 1000) - maximum characters per chunk
- `chunk_overlap` (default: 150) - characters repeated at boundaries

Scientific rationale: Useful as a baseline in ablation studies. Ensures a controlled chunk-size distribution for experiments where chunk-length variability would be a confound. Constraint enforced: `chunk_overlap` < `chunk_size`.

S2.1.2 Recursive Split (`strategy: "recursive"`) - *Default*

Hierarchical splitting that preserves semantic units by attempting separators in priority order: paragraph boundary (`"\n\n"`), line break (`"\n"`), word boundary (`" "`), and character level as a fallback.

S2.1.3 Paragraph Split (`strategy: "paragraph"`)

One document paragraph becomes one chunk, via Haystack's `DocumentSplitter(split_by="passage")`.

Parameters:

- `max_paragraph_length` (default: 2000) - warning threshold (not a hard truncation limit)

Scientific rationale: Optimal for structured documents (academic papers, legal texts) where paragraphs are deliberately written as semantic units. May degrade retrieval recall if very long paragraphs contain multiple distinct facts.

S2.1.4 Comparative Summary

The three chunking strategies exhibit different trade-offs in terms of semantic coherence, chunk-size regularity, and retrieval behaviour. As summarised in Table S2, recursive splitting provides the best balance for general-purpose retrieval scenarios, whereas paragraph-based chunking maximises semantic integrity in highly structured documents.

Property	Character	Recursive	Paragraph
Length uniformity	High	Medium	Low
Semantic coherence	Low	High	Very high
Sentence fragmentation	Common	Rare	None
Overlap support	Yes	Yes	No
Best for	Baselines	General use	Structured docs

Table S2: Comparison of chunking strategies available in RAGtio.

S2.2 Embedding Models

Engine: FastEmbed (ONNX runtime, CPU by default; no GPU required).
Source: HuggingFace model hub (downloaded once, cached locally).

S2.2.1 Dense Embeddings

Converts text to continuous vector representations:

$$\text{text} \xrightarrow{\text{tokenise}} \text{encoder layers} \xrightarrow{\text{pooling}} \text{L2-normalise} \longrightarrow \mathbf{v} \in \mathbb{R}^d$$

Model	Language	Dim	Notes
`intfloat/multilingual-e5-large`	Multi	1024	Default; instruction-tuned
`intfloat/multilingual-e5-base`	Multi	768	Lighter alternative
`BAAI/bge-m3`	Multi	1024	State-of-the-art multi-function
`BAAI/bge-small-en-v1.5`	EN	384	Efficient, English-only
`sentence-transformers/all-MiniLM-L6-v2`	EN	384	Fast, general purpose

Table S3: Supported dense embedding models with native dimension resolution.

S2.2.2 Sparse Embeddings (BM25-compatible)

Computes token-level importance weights for exact-match retrieval:

$$\text{text} \xrightarrow{\text{tokenise}} \text{BM25 TF-IDF weighting} \longrightarrow \{t_i : w_i\} \quad \text{(sparse vector)}$$

Implementation: `FastembedSparseDocumentEmbedder`. Sparse embeddings are always computed during indexing regardless of the configured retrieval mode, enabling future mode switching without reindexing.

S3 Retrieval Pipeline

Source: `app/retrieval.py`
Purpose: Given a query, return the most relevant document chunks.
Pattern: Lazy-initialised, cached pipeline per `(mode, collection)` pair.

S3.1 Retrieval Modes

S3.1.1 Dense Retrieval (`mode: "dense"`)

Approximate nearest-neighbour search in the dense embedding space:

$$q_{\text{text}} \xrightarrow{\text{FastembedTextEmbedder}} \mathbf{q} \in \mathbb{R}^d \xrightarrow{\text{HNSW search}} \text{Top-}K \text{ by cosine similarity}$$

Index structure: HNSW (Hierarchical Navigable Small World graph).
Similarity: Cosine similarity (default for L2-normalised vectors in Qdrant).

- \+ Semantic similarity beyond lexical overlap
- \+ Effective for paraphrasing and conceptual queries
- – May miss exact keyword matches
- – Sensitive to domain shift (OOV terms poorly represented)

S3.1.2 Sparse Retrieval (`mode: "sparse"`)

Inverted-index-style matching on BM25-weighted token vectors:

$$q_{\text{text}} \xrightarrow{\text{FastembedSparseTextEmbedder}} \{t_i : w_i\} \xrightarrow{\text{dot-product}} \text{Top-}K \text{ by BM25 score}$$

- \+ Exact term matching; interpretable scores
- \+ Effective for technical jargon, product names, codes
- – No semantic generalisation; fails on synonyms

S3.1.3 Hybrid Retrieval (`mode: "hybrid"`) - *Default*

Combines dense and sparse signals via Reciprocal Rank Fusion (RRF).

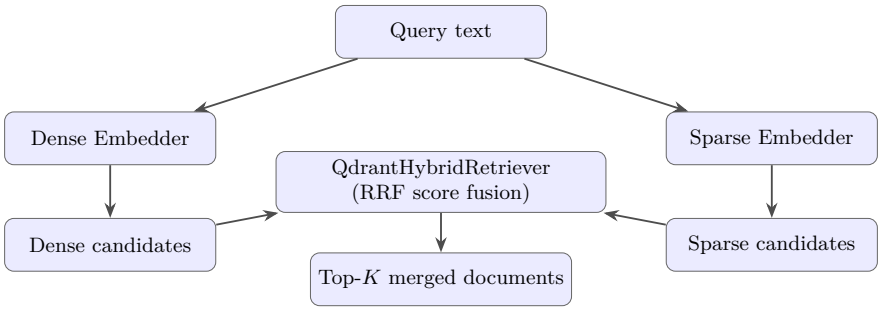

Figure S1: Hybrid retrieval pipeline.

Reciprocal Rank Fusion (RRF) merges the two ranked lists:

$$\text{RRF}(d) = \sum_{r \in \mathcal{R}} \frac{1}{k + \text{rank}_r(d)} \tag{1}$$

where $\mathcal{R}$ is the set of rankers (dense, sparse) and $k = 60$ is a smoothing constant [1].

+ Best overall recall in standard benchmarks
+ Robust across query types (semantic and lexical)
− Higher latency (two embedding passes, two ANN searches)

S3.2 Optional Reranking

```
reranker:
  enabled: false
  model: "cross-encoder/ms-marco-MiniLM-L-6-v2"
  batch_size: 16
  cache_dir: "/app/models/reranker"
```

Listing 1: Reranker configuration.

Architecture: Cross-encoder (`TransformersSimilarityRanker`).

Unlike bi-encoders [4] used for initial retrieval, a cross-encoder processes the query–document pair *jointly*:

$$[\texttt{CLS}]\ q\ [\texttt{SEP}]\ d\ [\texttt{SEP}] \xrightarrow{\text{Transformer}} h_{\texttt{CLS}} \xrightarrow{W \in \mathbb{R}^{1\times d}} s = \sigma(W h_{\texttt{CLS}}) \in [0,1]$$

Retrieve-then-rerank paradigm [3]:

$$\underbrace{\text{Query} \longrightarrow \text{Retrieval}(\text{top\_k} = 10)}_{\text{high recall, low cost}} \longrightarrow \underbrace{\text{Cross-Encoder} \longrightarrow \text{Top-}N(\text{top\_n\_after\_rerank} = 5)}_{\text{high precision, higher cost}}$$

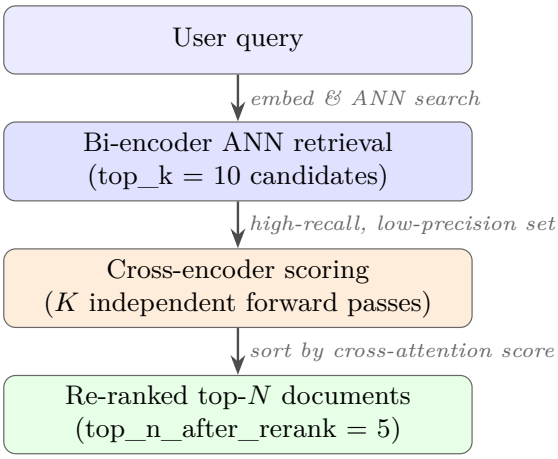

Figure S2: Two-stage retrieve-then-rerank pipeline. Stage 1 maximises recall at low cost; Stage 2 maximises precision using joint query–document encoding.

Bi-encoder vs. Cross-encoder comparison:

Property	**Bi-encoder (retrieval)**	**Cross-encoder (reranking)**
Query encoding	Independent of documents	Joint with each document
Document encoding	Pre-computed & cached	On-the-fly per query–doc pair
Attention	None between q and d	Full cross-attention
Inference cost	$O(1)$ + ANN search	$O(K)$ forward passes
Relevance precision	Moderate	High
Scalability	Millions of documents	Limited by K (~10–50)

Table S4: Architectural comparison of bi-encoder retrieval and cross-encoder reranking.

Cross-encoders achieve significantly higher relevance precision than bi-encoders owing to full cross-attention between query and document tokens, at the cost of $O(K)$ inference passes. The two-stage design bounds this cost: the cross-encoder never processes more than `top_k` candidates regardless of collection size.

Implementation: The reranker is loaded once (warm-up at first request) and held as a module-level singleton `_reranker`, ensuring that model weights are not reloaded on every query.

Model: `cross-encoder/ms-marco-MiniLM-L-6-v2` (6-layer MiniLM, 22M parameters, trained on MS MARCO passage ranking).

S4 RAG Q&A Pipeline

Source: `app/rag_qa.py`
Purpose: End-to-end question answering combining retrieval with LLM generation.
Modes: Synchronous (full response) and streaming (Server-Sent Events, token-by-token).

S4.1 Pipeline Stages

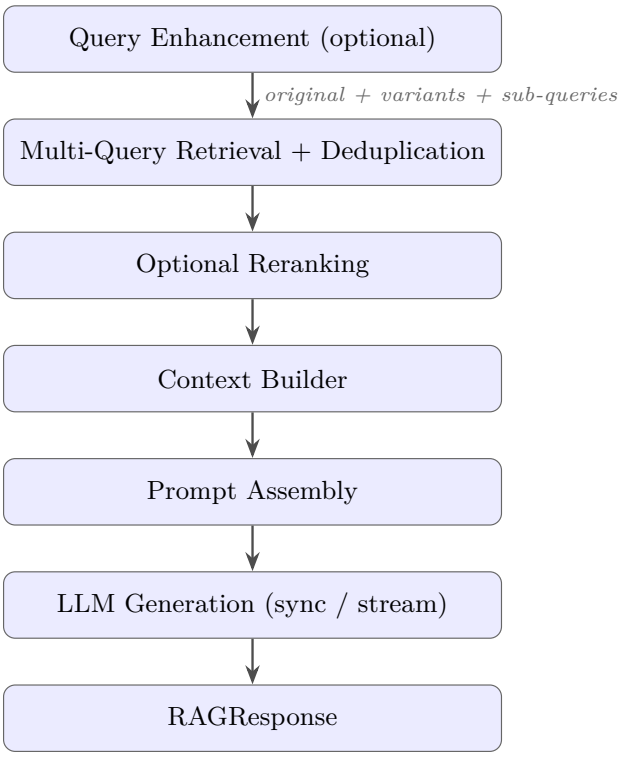

Figure S3: RAG Q&A pipeline data flow.

S4.2 Prompt Engineering

The system uses a two-message chat format supported by both LLM backends:

- **System message** (`llm.system_prompt`): role definition and grounding instructions.
- **User message** (`llm.rag_prompt_template`): context block (`{context}`) followed by the user's original question (`{question}`).

Both prompts are fully configurable in `config.yaml`.

S4.3 Response Data Model

```python
@dataclass
class RAGResponse:
    answer: str
    sources: List[SourceDocument]   # retrieved chunks with metadata
    query: str                       # original user query
    sub_queries: List[str]           # enhancement variants used
    retrieval_mode: str              # "dense" | "sparse" | "hybrid"
    n_docs_retrieved: int
    elapsed_time_seconds: float

@dataclass
class SourceDocument:
    content: str
    score: float                # retrieval / reranking score
    source: str                 # file path or URL
    page: Optional[int]         # PDF page number
    metadata: Dict[str, Any]
```

Listing 2: RAGResponse and SourceDocument dataclasses.

S5 Evaluation Pipeline

Sources: `app/evaluation.py`, `app/metrics.py`
Purpose: Quantitatively assess retrieval quality via information retrieval metrics.
Modes: Mode A (synthetic query generation) and Mode B (custom labelled dataset).

S5.1 Mode A - Synthetic Evaluation

Use case: No labelled dataset available; automated evaluation for development and regression testing.

Listing 3: Mode A evaluation algorithm (pseudo-code).

```
for i in 1..n_samples:
    chunk_i  = random_sample(qdrant_collection)
    query_i  = llm_generate_query(chunk_i, query_type)
    results_i = retrieve(query_i, top_k=max(k_values))
    metrics_i = compute_ir_metrics(results_i, relevant=[chunk_i.id])

aggregate = mean over all samples
```

Query generation modes (`query_type` parameter):

Mode	Strategy	Example output
`question`	Force question generation	"What is the annual revenue of X?"
`keywords`	Force keyword extraction	"annual revenue X financial report"
`auto`	Adaptive (retrieval-aware)	keywords for sparse; question for dense/hybrid

Table S5: Query generation modes for synthetic evaluation.

Rationale for `auto` *mode:* Sparse retrieval is optimised for keyword matching, so keyword queries are more appropriate probes. Dense and hybrid retrievers handle natural language, so question-style queries better stress-test semantic understanding.

S5.1.1 Query Generation for Mode A

The `_resolve_query_kind()` function selects the generation strategy at runtime based on the active retrieval mode:

```python
def _resolve_query_kind(cfg: AppConfig) -> QueryKind:
    query_type = cfg.evaluation.mode_a.query_type
    if query_type == "auto":
        return "keywords" if cfg.retrieval.mode == "sparse" else "question"
    return query_type   # "question" or "keywords" forced by config
```

Listing 4: Adaptive query kind resolution.

Question generation prompt (used for `dense` and `hybrid`):

```
Read the following text and formulate a question that the text directly answers.
Return only the question, without any additional text.
Text: {chunk}
```

Keyword generation prompt (used for `sparse`):

```
Read the following text and extract 3 to 6 keywords or short key phrases that represent it.
Return only the keywords separated by commas, without any additional text.
Text: {chunk}
```

query_type	Retrieval mode	Strategy	Rationale
auto	sparse	keywords	Matches BM25 lexical signal
auto	dense/hybrid	question	Tests semantic understanding
question	any	question	Forces NL query regardless of mode
keywords	any	keywords	Forces keyword query regardless of mode

Table S6: Query generation strategy by `query_type` setting and retrieval mode.

Scientific rationale: BM25-based sparse retrieval computes scores via token-level overlap weighted by term frequency and inverse document frequency. Natural-language questions contain high-frequency stopwords (articles, prepositions, auxiliary verbs) that dilute BM25 scores for the informative content terms. Keyword queries, conversely, concentrate weight on discriminative terms and directly probe the lexical matching capability of sparse retrieval. For dense and hybrid retrievers, question-style probes are more appropriate because the bi-encoder is trained to align *question↔passage* pairs (as in the MS MARCO and Natural Questions training regimes), and thus question queries better exercise the model's semantic generalisation.

Both prompt templates and their defaults are fully configurable via `config.yaml` under `evaluation.mode_a`.

S5.2 Mode B - Custom Dataset Evaluation

Use case: Labelled dataset available; rigorous offline evaluation.

Input format (JSON array):

```
[
  {
    "query": "What is the capital of France?",
    "relevant_ids": ["chunk-abc-123", "chunk-def-456"],
    "expected_text": "optional reference answer"
  },
  ...
]
```

Listing 5: Mode B evaluation algorithm.

```
for each (query, relevant_ids) in dataset:
    results       = retrieve(query, top_k=max(k_values))
    retrieved_ids = [r.id for r in results]
```

```
    metrics       = compute_ir_metrics(retrieved_ids, relevant_ids, k_values)

aggregate = mean over all queries
```

S5.3 Evaluation Report

Both modes produce an `EvaluationReport` dataclass serialised to JSON:

```python
@dataclass
class EvaluationReport:
    mode: Literal["A", "B"]
    n_samples: int
    ir_metrics: IRMetrics
    gen_metrics: Optional[GenMetrics]  # reserved; currently None
    k_values: List[int]
    retrieval_mode: str                # dense | sparse | hybrid
    reranker_used: bool
    elapsed_time_seconds: float
    per_sample_results: List[dict]     # per-query breakdown

@dataclass
class IRMetrics:
    mrr: float
    recall_at_k:   Dict[int, float]   # e.g. {1:0.45, 3:0.72, 5:0.81, 10:0.91}
    hit_rate_at_k: Dict[int, float]
    ndcg_at_k:     Dict[int, float]
```

Listing 6: EvaluationReport and IRMetrics dataclasses.

S5.4 Asynchronous Job Management

Evaluation runs are submitted as background jobs:

```
POST /api/eval      -> {"job_id": "eval-uuid-123", "status": "running"}
GET  /api/eval/{id} -> {"status": "done",  "report": {...}}
                    -> {"status": "error", "message": "..."}
```

In-memory store: `_eval_jobs: Dict[str, JobState]`. Reports are also persisted to `evaluation.output_dir` (default: `/app/eval_reports`).

S6 Information Retrieval Metrics

Source: `app/metrics.py`

All metrics are computed on ranked lists of retrieved chunk IDs against a ground-truth set of relevant chunk IDs.

S6.1 Recall@K

Definition: Fraction of relevant documents found in the top-K results.

$$\text{Recall@K} = \frac{|\{\text{retrieved}_{1:K}\} \cap \{\text{relevant}\}|}{|\text{relevant}|} \tag{2}$$

Range: $[0, 1]$; higher is better.

Note: In Mode A evaluation $|\text{relevant}| = 1$ (single ground-truth chunk), so Recall@K $\equiv$ HitRate@K in that setting.

```python
def recall_at_k(retrieved_ids, relevant_ids, k):
    return len(set(retrieved_ids[:k]) & set(relevant_ids)) / len(relevant_ids)
```

S6.2 Hit Rate@K

Definition: Binary indicator; 1 if at least one relevant document appears in the top-K results.

$$\text{HitRate@K} = \mathbb{1}[|\{\text{retrieved}_{1:K}\} \cap \{\text{relevant}\}| > 0] \tag{3}$$

Averaged across queries to obtain a probability.

```
def hit_rate_at_k(retrieved_ids, relevant_ids, k):
    return float(len(set(retrieved_ids[:k]) & set(relevant_ids)) > 0)
```

S6.3 Mean Reciprocal Rank (MRR)

Definition: Reciprocal of the rank of the first relevant document, averaged over all queries.

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q} \tag{4}$$

where rank_q is the position of the first relevant result for query q (0 if not found).

Range: $[0, 1]$; higher is better. MRR $= 0.5$ implies the first relevant document is at rank 2 on average.

```
def reciprocal_rank(retrieved_ids, relevant_ids):
    for i, doc_id in itemize(retrieved_ids):
        if doc_id in relevant_ids:
            return 1.0 / (i + 1)
    return 0.0
```

Scientific note: MRR is most appropriate for navigational queries, where users inspect results in order and stop at the first satisfactory item. It heavily penalises systems that rank relevant documents low.

S6.4 Normalised Discounted Cumulative Gain (NDCG@K)

Definition: Position-aware metric that rewards ranking relevant documents higher.

$$\text{DCG@K} = \sum_{i=1}^{K} \frac{\text{rel}_i}{\log_2(i+1)} \tag{5}$$

$$\text{IDCG@K} = \sum_{i=1}^{\min(|\text{relevant}|, K)} \frac{1}{\log_2(i+1)} \tag{6}$$

$$\text{NDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}} \tag{7}$$

where $\text{rel}_i \in \{0, 1\}$ (binary relevance).

Range: $[0, 1]$; NDCG@K $= 1$ iff all relevant documents appear at the top positions.

```
def ndcg_at_k(retrieved_ids, relevant_ids, k):
    dcg = sum(
        1.0 / math.log2(i + 2)
        for i, doc_id in itemize(retrieved_ids[:k])
        if doc_id in relevant_ids
    )
    idcg = sum(
        1.0 / math.log2(i + 2)
        for i in range(min(len(relevant_ids), k))
    )
    return dcg / idcg if idcg > 0 else 0.0
```

Scientific note: NDCG is the standard metric for Learning-to-Rank evaluation [2]. Unlike MRR, it accounts for multiple relevant documents at various positions.

S6.5 Metric Selection Guidance

Metric	Best for	Sensitive to
Recall@K	Coverage, multi-evidence tasks	Total number of relevant docs
HitRate@K	Any-relevant-found tasks	First relevant doc presence
MRR	First-result quality, navigational	Rank of first relevant doc
NDCG@K	Multi-relevant ranking quality	All relevant doc positions

Table S7: Metric selection guide for RAG evaluation.

Recommended primary metric for RAG evaluation: NDCG@5 or NDCG@10, as it captures both coverage and ranking quality across the documents that will actually be passed to the LLM context window.

S7 Case Study results

This supplementary section reports the complete per-query retrieval results for all four biomedical domains. Aggregate cross-domain results are presented in the Quantitative Results subsection of the main manuscript, while the following tables provide the detailed query-level breakdown for each domain.

S8 Oncology domain

Evaluation outcomes for the Oncology domain, following the same order as in the main text.

S8.1 Mode A - Per-Query Results

Table S8 reports the individual retrieval outcomes for all 41 synthetic queries evaluated in Mode A. Queries are grouped by outcome: 25 achieve perfect retrieval (MRR = 1.0), 5 retrieve the correct chunk at a rank greater than one (partial retrieval, $0 < \text{MRR} < 1$), and 11 fail to surface the expected chunk within the top-10 results. The aggregate results across all domains are summarised in the Mode A aggregate-results table of the main manuscript.

#	Query	MRR	Recall@5	Recall@10
	Perfect retrieval (MRR = 1.0)			
1	Which genetic and cytogenetic alterations place AML in the ELN 2022 adverse prognostic group?	1.000	1.000	1.000
2	How can combining viruses with CD19-targeted CAR-T cell therapy achieve homogenous expression of tumor antigens and overcome tumor antigen escape?	1.000	1.000	1.000
7	What are the five types of MDSC-targeted therapies?	1.000	1.000	1.000
8	In what years were CLL and B-ALL associated with CAR-T cell therapy?	1.000	1.000	1.000
9	Do the authors have any competing interests?	1.000	1.000	1.000
10	What are the recent studies on bispecific T-cell engagers for relapsed/refractory acute myeloid leukemia?	1.000	1.000	1.000
13	What is necessary to draw definitive conclusions regarding the risks associated with CAR-T cell therapy?	1.000	1.000	1.000
14	Who published the first report of CD19 CAR-T therapy in a patient?	1.000	1.000	1.000
15	What cytokine has been identified as key for enhancing NK-cell activity?	1.000	1.000	1.000
17	What immunosuppressive features are associated with AML?	1.000	1.000	1.000
19	What is the biggest problem facing cancer immunotherapy?	1.000	1.000	1.000
20	What immune checkpoint molecules mediate inhibition of tumor-specific T cells?	1.000	1.000	1.000
22	What tumor vaccine has been approved by the US FDA for treating advanced prostate cancer?	1.000	1.000	1.000
23	What do LSCs stand for?	1.000	1.000	1.000
26	What was the overall response rate (ORR) for tisa-cel in pediatric patients with B-ALL who were refractory to standard chemotherapy or relapsed after HSCT?	1.000	1.000	1.000
27	What was the response rate when monalizumab was combined with cetuximab for recurrent or metastatic head and neck squamous cell carcinoma?	1.000	1.000	1.000
29	What are potential tumor targets in myeloid leukemia?	1.000	1.000	1.000

#	Query	MRR	Recall@5	Recall@10
30	How does interferon gamma contribute to immune escape in tumor cells?	1.000	1.000	1.000
32	What strategies are being evaluated to permit physicians to modulate CAR activity based on clinical needs?	1.000	1.000	1.000
33	What are the limitations of immune checkpoint therapy?	1.000	1.000	1.000
35	Why did the FDA grant tissue-agnostic approval to pembrolizumab in 2017?	1.000	1.000	1.000
36	What is T-VEC?	1.000	1.000	1.000
37	Why do PD-L1 testing results vary between clinical trials?	1.000	1.000	1.000
39	What does CAR-T cell therapy involve?	1.000	1.000	1.000
41	How many clinical trials of CAR-T cell therapy for AML are currently registered?	1.000	1.000	1.000
	Partial retrieval ($0 < MRR < 1$)			
3	What are potential targets for CAR T-cell therapy in acute myeloid leukemia?	0.500	1.000	1.000
4	What are the NCT numbers for clinical trials involving idecabtagene vicleucel in multiple myeloma?	0.200	1.000	1.000
16	What strategy might be used to overcome T cell exhaustion and enhance CAR-T cell efficacy in AML treatment?	0.250	1.000	1.000
25	What year was the review on immune checkpoint inhibitors by Kroemer and Zitvogel published?	0.500	1.000	1.000
28	What did Dr. Michel Sadelain improve in CAR-T cell design to enhance persistence and survival?	0.500	1.000	1.000
	Failed retrieval ($MRR = 0$)			
5	What clinical trials are referenced in this text?	0.000	0.000	0.000
6	What are the main challenges in extending CAR-T therapy to solid tumors?	0.000	0.000	0.000
11	What is the overall response rate (ORR) for tisa-cel versus standard of care (SOC) in adult patients with aggressive B-NHL?	0.000	0.000	0.000
12	What are the clinical applications of the biomarkers listed?	0.000	0.000	0.000
18	What are examples of novel immunotherapeutic agents in early-phase clinical trials?	0.000	0.000	0.000
21	How do STAR-T cells compare to conventional CAR-T strategies in low neoantigen density contexts?	0.000	0.000	0.000
24	What are the clinical efficacy and safety profiles of the non-immune checkpoint blockade agents listed in the table?	0.000	0.000	0.000
31	What are the four sources from which ctDNA fragments are derived?	0.000	0.000	0.000
34	What is the purpose of the ZUMA-25 clinical trial?	0.000	0.000	0.000
38	What was the overall response rate in the LEGEND-2 trial?	0.000	0.000	0.000
40	What is the target antigen for the CAR-T cell therapies described in these clinical trials?	0.000	0.000	0.000

Table S8: Mode A per-query results for the Oncology domain (hybrid retrieval with cross-encoder re-ranking). Queries are grouped by retrieval outcome.

Analysis of failure cases. The 11 failed queries share a common characteristic: they require synthesising information that is either scattered across multiple disconnected passages (e.g., Q5, Q12, Q24) or embedded within large structured tables (e.g., Q11, Q34, Q38, Q40). In such cases, the expected chunk is a specific

table row or a cross-document summary that does not form a self-contained semantic unit, making it difficult for the dense retriever to match the query intent. The 5 partially retrieved queries all succeed at Recall@10, indicating that the relevant content is present in the corpus but is outranked by semantically similar yet non-target passages at early positions.

S8.2 Mode B - Per-Query Results

Table S9 provides the per-query breakdown for the five manually annotated queries in Mode B. The number of relevant chunks manually annotated varies substantially across queries, ranging from 4 to 13, which directly affects recall values.

Table S9: Mode B per-query results for the Oncology domain (hybrid retrieval with cross-encoder re-ranking). $|\mathcal{R}|$ denotes the number of manually annotated relevant chunks.

Query ID	$\|\mathcal{R}\|$	MRR	R@1	R@5	R@10	nDCG@5
Q1	13	1.000	0.077	0.154	0.154	0.470
Q2	9	0.000	0.000	0.000	0.000	0.000
Q3	12	1.000	0.083	0.250	0.250	0.655
Q4	4	1.000	0.250	0.500	0.500	0.637
Q5	4	1.000	0.250	1.000	1.000	0.905

The five queries cited in Table S9 are listed below:

Q1: *Which is the role of monoclonal antibodies?*

Q2: *Do CAR-T cells play a role in cancer treatment?*

Q3: *Which is the role of dendritic cells in tumour immunotherapy?*

Q4: *What is immunotherapy resistance?*

Q5: *What is precision oncology and how it works?*

Query 2 (Q2) achieves zero retrieval across all cutoffs.This query is deliberately broad: the manually identified 9 relevant chunks spanning multiple documents, but the dense retriever fails to match the query's high-level intent to any single chunk, and the cross-encoder re-ranker cannot recover from the initial miss. In contrast, Query 5 (Q5) achieves perfect Recall@5 and Recall@10, as its 4 relevant chunks are concentrated in fewer documents and are more topically focused. The aggregate results across all domains are summarised in the Mode B aggregate-results table of the main manuscript.

S9 Diabetes domain

Full evaluation metrics and qualitative analyses for the Diabetes corpus.

S9.1 Mode A - Per-Query Results

Table S10 reports the individual retrieval outcomes for all 40 synthetic queries evaluated in Mode A. Queries are grouped by outcome: 20 achieve perfect retrieval (MRR = 1.0), 15 retrieve the correct chunk at a rank greater than one ($0 < \text{MRR} < 1$), and 5 fail to surface the expected chunk within the top-10 results. The aggregate results across all domains are summarised in the Mode A aggregate-results table of the main manuscript.

#	Query	MRR	Recall@5	Recall@10
	Perfect retrieval (MRR = 1.0)			
4	What predicts cardiovascular outcomes in type 2 diabetic patients?	1.000	1.000	1.000

#	Query	MRR	Recall@5	Recall@10
5	Do fatty acids desensitize insulin stimulation of glucose uptake in muscle cells?	1.000	1.000	1.000
7	What is the aim of the ANDIS study?	1.000	1.000	1.000
8	Why does adopting a tailored therapeutic approach addressing overall patient's comorbid conditions become essential?	1.000	1.000	1.000
9	How much less effective are semaglutide and tirzepatide in preobese and obese people with T2D compared to those without diabetes?	1.000	1.000	1.000
10	What is the effect of hyperinsulinemia on diabetes progression?	1.000	1.000	1.000
13	Does the model developed here explain disease progression mechanistically or assess the current state?	1.000	1.000	1.000
17	Can researchers better map the aetiological processes that contribute to diabetes risk?	1.000	1.000	1.000
19	What unifying term has been defined for the coexistence of these conditions?	1.000	1.000	1.000
21	Is epicardial adipose tissue an emerging biomarker of cardiovascular complications in type 2 diabetes?	1.000	1.000	1.000
23	What are the target genes activated by NF-κB?	1.000	1.000	1.000
27	In an insulin-resistant state, how do kinases inhibit insulin action?	1.000	1.000	1.000
28	What are the cardiorenal mechanisms of action of glucagon-like-peptide-1 receptor agonists?	1.000	1.000	1.000
31	What are the key references on adipose tissue inflammation and metabolic disease?	1.000	1.000	1.000
33	How can the slope of trajectories from multiple OGTTs spaced in time indicate a patient's position on the progression path?	1.000	1.000	1.000
35	What is the final shared pathway for progressive renal impairment in diabetic kidney disease?	1.000	1.000	1.000
37	What effect does decreasing the parameter hepaSI in Equation S4 have on HGP at a fixed SI?	1.000	1.000	1.000
38	Which cluster of individuals with type 2 diabetes responds better to thiazolidinediones?	1.000	1.000	1.000
39	How does intensive glycemic control affect the progression of DPN and CAN in T2D?	1.000	1.000	1.000
40	How does the risk for cardiovascular events and death change with declining estimated GFR?	1.000	1.000	1.000
	Partial retrieval ($0 < MRR < 1$)			
1	Which institutions and departments are the authors affiliated with?	0.250	1.000	1.000
6	Is the combined glucose impairment (CGI) state obligatory for progression from impaired glucose tolerance (IGT) to type 2 diabetes (T2D)?	0.500	1.000	1.000
11	What validated methods and critical reference values are lacking for salivary glucose testing?	0.200	1.000	1.000
12	What can diabetes be considered as, given the defects in physiological processes of tissue repair?	0.500	1.000	1.000
16	To what extent do heart failure and chronic kidney disease promote the development of type 2 diabetes or worsen its control?	0.500	1.000	1.000

#	Query	MRR	Recall@5	Recall@10
18	What is the overall T2D incidence rate in patients with CKD according to the CRIC Study?	0.500	1.000	1.000
20	Does correcting metabolic acidosis improve insulin resistance in chronic kidney disease?	0.500	1.000	1.000
24	Why do clinicians need to recognize that type 2 diabetes is multifaceted?	0.500	1.000	1.000
25	What is the sequence of glucose abnormalities in the progression from normal glucose tolerance to type 2 diabetes?	0.500	1.000	1.000
26	What does NF-κB activation in T2D inhibit?	0.500	1.000	1.000
29	What are the recent clinical guidelines for managing diabetes in chronic kidney disease?	0.250	1.000	1.000
30	What contributes to the decline in kidney function in DKD?	0.500	1.000	1.000
32	Which hypoglycemic agents have been proven to have very high efficacy?	0.250	1.000	1.000
34	What is the relationship between nonalcoholic fatty liver disease and the risk of cardiovascular disease?	0.500	1.000	1.000
36	What does this study focus on rather than assessment?	0.500	1.000	1.000
	Failed retrieval (MRR = 0)			
2	How do the HOMA-IR values change as the insulin resistance index increases?	0.000	0.000	0.000
3	What references are listed in this document?	0.000	0.000	0.000
14	What is type 2 diabetes currently considered to be?	0.000	0.000	0.000
15	Why have SGLT-2 inhibitors become the first choice of hypoglycemic agent for patients with T2D and HF or for those with multiple risk factors for HF?	0.000	0.000	0.000
22	What are the references cited in this document?	0.000	0.000	0.000

Table S10: Mode A per-query results for the Diabetes domain (hybrid retrieval with cross-encoder reranking). Queries are grouped by retrieval outcome.

Analysis of failure cases. The 5 failed queries fall into two categories: reference-list queries (Q3, Q22), which require enumerating all citations in a document and thus lack a single target chunk; and queries whose expected answer is embedded in a figure caption or a passage that does not form a self-contained semantic unit (Q2, Q14, Q15). Notably, 15 out of 40 queries (37.5%) achieve partial retrieval, a higher proportion than in the Oncology domain, reflecting the greater topical overlap among diabetes documents where multiple passages discuss similar mechanisms (e.g., insulin resistance, NF-κB signalling). All partially retrieved queries succeed at Recall@10, confirming that the relevant content is present but outranked at early positions.

S9.2 Mode B - Per-Query Results

Table S11 provides the per-query breakdown for the five manually annotated queries in Mode B. The number of relevant chunks manually annotated varies across queries, ranging from 1 to 4, which directly affects recall values.

The queries cited in Table S11 are listed below:

Q1: *What is the cardio-renal-metabolic connection?*

Q2: *Is diabetes a multifaceted disease?*

Q3: *What is the role of steroid hormones in diabetes?*

Q4: *Which signaling pathways are involved in diabetes?*

Table S11: Mode B per-query results for the Diabetes domain (hybrid retrieval with cross-encoder re-ranking). $|\mathcal{R}|$ denotes the number of manually annotated relevant chunks.

Query ID	$\|\mathcal{R}\|$	MRR	R@1	R@5	R@10	nDCG@5
Q1	4	1.000	0.250	0.750	0.750	0.832
Q2	2	1.000	0.500	0.500	0.500	0.613
Q3	1	0.000	0.000	0.000	0.000	0.000
Q4	2	1.000	0.500	0.500	0.500	0.613
Q5	2	0.500	0.000	0.500	0.500	0.387

Q5: *What is the IFG-First Pathway?*

Query 3 (Q3) achieves zero retrieval across all cutoffs. This query targets a single relevant chunk discussing steroid hormones, which is not surfaced by the dense retriever within the top-10 results, likely because the topic is only tangentially addressed in the corpus and the chunk lacks strong semantic overlap with the query formulation. Query 1 (Q1) achieves the highest Recall@5 (0.750) among the five queries, as its 4 relevant chunks are concentrated within a single review article on the cardio-renal-metabolic connection. The aggregate results across all domains are summarised in the Mode B aggregate-results table of the main manuscript.

S10 Pharmacology domain

Performance of the retrieval and re-ranking pipeline on the Pharmacology domain.

S10.1 Mode A - Per-Query Results

Table S12 reports the individual retrieval outcomes for all 40 synthetic queries evaluated in Mode A. Queries are grouped by outcome: 28 achieve perfect retrieval (MRR = 1.0), 8 retrieve the correct chunk at a rank greater than one ($0 < \text{MRR} < 1$), and 4 fail to surface the expected chunk within the top-10 results. The aggregate results across all domains are summarised in the Mode A aggregate-results table of the main manuscript.

#	Query	MRR	Recall@5	Recall@10
	Perfect retrieval (MRR = 1.0)			
2	What are the predicted effects of antibiotics and antifungals on DOACs exposure and pharmacological activity?	1.000	1.000	1.000
3	How many missense variants for CYP2C9 and CYP2C19 have been functionally validated compared to those listed in the database?	1.000	1.000	1.000
4	Who received financial support from EDRA S.p.A. and who are employees of EDRA S.p.A.?	1.000	1.000	1.000
5	How does the interaction between dabigatran and verapamil depend on the time of administration and the formulation of verapamil?	1.000	1.000	1.000
6	Why should the evaluation of PD DDIs not be overlooked in clinical drug development?	1.000	1.000	1.000
7	What is the likely true incidence of adverse drug reactions to anaesthetic drugs?	1.000	1.000	1.000
8	Do fluvastatin, fenofibrate, gemfibrozil, ezetimibe, and PCSK9 inhibitors have clinically relevant interactions with DOACs?	1.000	1.000	1.000

#	Query	MRR	Recall@5	Recall@10
9	What does simulation result in for patient response to drug therapies?	1.000	1.000	1.000
10	How can one determine whether a drug combination is Loewe synergistic or antagonistic?	1.000	1.000	1.000
11	What enabled and stimulated pharmacogenomic discovery and clinical implementation in the United States?	1.000	1.000	1.000
13	What does the multiscale network model identify in pancreatic cancer cells?	1.000	1.000	1.000
14	What do NGS, SNP, and GWAS stand for?	1.000	1.000	1.000
16	When should drug challenging be considered for suspected adverse drug reactions?	1.000	1.000	1.000
17	By what percentage do PPIs reduce dabigatran's AUC and Cmax?	1.000	1.000	1.000
20	Which drugs used in the treatment of COVID-19 are contraindicated with DOACs?	1.000	1.000	1.000
21	What happens to aspirin's inhibition of platelet aggregation when ibuprofen is administered daily with aspirin?	1.000	1.000	1.000
22	What do the examples illustrate in pharmacogenetics?	1.000	1.000	1.000
24	What are the FDA pharmacogenomic biomarkers for non-cancer therapeutics?	1.000	1.000	1.000
25	Which DOAC requires dose adjustment when co-administered with erythromycin?	1.000	1.000	1.000
26	Why will pharmacogenomics see the earliest and broadest clinical implementation?	1.000	1.000	1.000
28	Why did Weinshilboum and Wang's work become relevant at the beginning of the 21st century?	1.000	1.000	1.000
29	Does co-administration of amiodarone with dabigatran increase the risk of major bleeding?	1.000	1.000	1.000
30	What aspects of pharmacogenomics will this review survey?	1.000	1.000	1.000
31	What were the plasma concentrations of edoxaban with and without amiodarone?	1.000	1.000	1.000
34	What is the term for moving beyond pharmacogenomics to integrate genomics with other omic data?	1.000	1.000	1.000
36	How were the included studies stratified by age groups and study settings?	1.000	1.000	1.000
37	What factors contribute to adverse drug reactions?	1.000	1.000	1.000
39	Does the fixed-dose combination of sofosbuvir and ledipasvir have potential drug interactions with DOACs?	1.000	1.000	1.000
	Partial retrieval ($0 < MRR < 1$)			
1	What is the most commonly used definition of an adverse drug reaction (ADR)?	0.333	1.000	1.000
12	What are the definitions of synergistic, additive, and antagonistic pharmacokinetic drug-drug interactions?	0.500	1.000	1.000
18	Which statins are safer alternatives to simvastatin for patients taking DOACs?	0.500	1.000	1.000
19	When might the Bliss independence method predict an outcome for the same drug at different concentrations?	0.500	1.000	1.000
27	Is co-administration of dronedarone and dabigatran contraindicated?	0.200	1.000	1.000
33	What percentage of studies included specific details about the list of conditions used to define multimorbidity?	0.250	1.000	1.000

#	Query	MRR	Recall@5	Recall@10
38	Do NSAIDs increase the risk of bleeding when used with anticoagulants?	0.333	1.000	1.000
40	What is the oral bioavailability of dabigatran etexilate (Pradaxa®) after co-medication with pantoprazole?	0.500	1.000	1.000
	Failed retrieval (MRR = 0)			
15	What does this model establish as a quantitative foundation for?	0.000	0.000	0.000
23	When was the article downloaded?	0.000	0.000	0.000
32	What references are cited in this section?	0.000	0.000	0.000
35	What is one of the primary challenges in the assessment of PD DDIs?	0.000	0.000	0.000

Table S12: Mode A per-query results for the Pharmacology domain (hybrid retrieval with cross-encoder re-ranking). Queries are grouped by retrieval outcome.

Analysis of failure cases. The 4 failed queries include a reference-list query (Q32), a metadata query about article download date (Q23), and two queries whose expected answers are embedded in passages that lack a self-contained semantic unit (Q15, Q35). The Pharmacology domain achieves the highest perfect-retrieval rate among all domains (70%), likely reflecting the more structured and technical nature of pharmacological documents, where key facts (e.g., drug interaction tables, pharmacogenomic definitions) are concentrated in well-defined passages. All 8 partially retrieved queries succeed at Recall@10, confirming that the relevant content is present but outranked at early positions.

S10.2 Mode B - Per-Query Results

Table S13 provides the per-query breakdown for the five manually annotated queries in Mode B. The number of relevant chunks manually annotated varies across queries, ranging from 1 to 6, which directly affects recall values.

Table S13: Mode B per-query results for the Pharmacology domain (hybrid retrieval with cross-encoder re-ranking). $|\mathcal{R}|$ denotes the number of manually annotated relevant chunks.

Query ID	$\lvert\mathcal{R}\rvert$	MRR	R@1	R@5	R@10	nDCG@5
Q1	5	0.200	0.000	0.200	0.200	0.131
Q2	5	1.000	0.200	0.600	0.600	0.699
Q3	4	0.000	0.000	0.000	0.000	0.000
Q4	1	0.500	0.000	1.000	1.000	0.631
Q5	6	1.000	0.167	0.333	0.333	0.485

The queries cited in Table S13 are listed below:

Q1: *Which are the reactions to anaesthetic drugs?*

Q2: *Why is dronedarone dangerous?*

Q3: *What is the difference between a synergistic and an additive drug interaction?*

Q4: *What are the treatments of antimicrobial infections?*

Q5: *Why are disease-drug interactions distinct from drug-drug interactions, and why is this distinction particularly important in patients with multimorbidity?*

Query 3 (Q3) achieves zero retrieval across all cutoffs. This query targets 4 relevant chunks discussing synergistic vs. additive drug interactions, but the dense retriever fails to match the query's conceptual

framing to any single chunk, as the relevant content is distributed across multiple documents with different terminological conventions. Query 5 (Q5), despite having the largest relevant set ($|\mathcal{R}|$=6), achieves the highest Hit Rate at rank 1 (MRR = 1.0), as its relevant chunks are concentrated in a single review article on multimorbidity and polypharmacy. The aggregate results across all domains are summarised in the Mode B aggregate-results table of the main manuscript.

S11 Infectious Diseases domain

S11.1 Mode A - Per-Query Results

Table S14 reports the individual retrieval outcomes for all 38 synthetic queries evaluated in Mode A. Queries are grouped by outcome: 24 achieve perfect retrieval (MRR = 1.0), 10 retrieve the correct chunk at a rank greater than one (0 < MRR < 1), and 4 fail to surface the expected chunk within the top-10 results. The aggregate results across all domains are summarised in the Mode A aggregate-results table of the main manuscript.

#	Query	MRR	Recall@5	Recall@10
	Perfect retrieval (MRR = 1.0)			
3	What factors predict acute graft-versus-host disease?	1.000	1.000	1.000
4	Where is supplementary material available?	1.000	1.000	1.000
7	What are the important consequences of transboundary disease spread through legal and illegal trade of live animals for global biodiversity?	1.000	1.000	1.000
9	What are the key references on heterogeneity in mycobacteria?	1.000	1.000	1.000
10	What year was the study by Kaur et al. on olopatadine hydrochloride published?	1.000	1.000	1.000
11	Is there a consensus among medical professionals on an optimal salvage therapy regimen for relapsed or refractory HLH?	1.000	1.000	1.000
13	What can result from overproduction of cytokines in pneumonia?	1.000	1.000	1.000
14	Do neutrophils benefit or harm the host during M. tuberculosis infection?	1.000	1.000	1.000
16	What is the title of the study that investigated immune asynchrony in COVID-19 pathogenesis?	1.000	1.000	1.000
17	What factors should be considered when determining the need for genetic testing in patients with suspected HLH?	1.000	1.000	1.000
18	What is the new threat of infectious diseases likely to come from?	1.000	1.000	1.000
19	What factors have reduced mortality and morbidity from infectious diseases in the past two decades?	1.000	1.000	1.000
20	Why are the precise cellular mechanisms through which autophagy controls M. tuberculosis in macrophages ex vivo incompletely understood?	1.000	1.000	1.000
22	What was the global number of people affected by community-acquired urinary tract infections caused by ESKAPE pathogens and the associated death toll in 2019?	1.000	1.000	1.000
23	How many clinical trials investigating MSC treatment have been recorded in the NIH Clinical Trial Database?	1.000	1.000	1.000
27	Why has caseum emerged as a niche of particular interest in tuberculosis research?	1.000	1.000	1.000
30	What is El Ni no?	1.000	1.000	1.000

#	Query	MRR	Recall@5	Recall@10
31	What coagulation disorder can result from severe hepatic dysfunction?	1.000	1.000	1.000
32	What role do neutrophil extracellular traps play in linking innate and adaptive immunity?	1.000	1.000	1.000
33	How does interferon-gamma upregulate human angiotensinogen (AGT) gene transcription?	1.000	1.000	1.000
34	How many structurally novel molecules with potent activity against A. baumannii were identified?	1.000	1.000	1.000
35	What does activation of the NLRP3 inflammasome contribute to in COVID-19?	1.000	1.000	1.000
37	How many deaths are associated with bacterial antimicrobial resistance in 2019?	1.000	1.000	1.000
38	What causes bone marrow suppression in aGVHD?	1.000	1.000	1.000
	Partial retrieval ($0 < MRR < 1$)			
1	What immune changes occur in sepsis patients that are associated with poor prognosis?	0.500	1.000	1.000
2	When might TLR4 inhibitors benefit septic patients?	0.500	1.000	1.000
5	What role do T cells play in the pathogenesis of ICI-associated myocarditis?	0.500	1.000	1.000
6	Which organizations provided funding for this research?	0.500	1.000	1.000
12	What are recent studies on myocarditis?	0.500	1.000	1.000
15	How does Mycobacterium tuberculosis transmit between individuals?	0.200	1.000	1.000
21	What are the three categories of mechanisms of antibiotic resistance (AMR)?	0.500	1.000	1.000
25	What has been observed in animal studies regarding IL-6 and TNF levels in sepsis mice following $\beta 2$ microglobulin adsorption column treatment?	0.500	1.000	1.000
26	Is pulmonary disease the sole terminal stage for tuberculosis transmission?	0.500	1.000	1.000
29	What factors predict a higher cumulative incidence of grade 2–4 CRS or ICANS after CAR-T therapy?	0.333	1.000	1.000
	Failed retrieval ($MRR = 0$)			
8	What are the authors and their institutional affiliations for this publication?	0.000	0.000	0.000
24	What strategies are used to overcome cytokine storm in different diseases?	0.000	0.000	0.000
28	What clinical trials are described in this text?	0.000	0.000	0.000
36	What are the references listed in this text?	0.000	0.000	0.000

Table S14: Mode A per-query results for the Infectious Diseases domain (hybrid retrieval with cross-encoder re-ranking). Queries are grouped by retrieval outcome.

Analysis of failure cases. The 4 failed queries include two reference-list queries (Q28, Q36), an author-affiliation query (Q8), and a broad thematic query about cytokine storm strategies (Q24). Reference-list and author-affiliation queries consistently fail across domains, as the relevant content is typically formatted as structured metadata (lists, tables) that does not match the semantic representation of the query. Query 24 targets a section heading ("Overcoming strategies of CS in different diseases") whose content spans multiple subsections, making it difficult for the retriever to map the query to a single chunk. All 10 partially retrieved queries succeed at Recall@10, confirming that the relevant content is present but outranked at early positions.

S11.2 Mode B - Per-Query Results

Table S15 provides the per-query breakdown for the five manually annotated queries in Mode B. The number of relevant chunks manually annotated varies across queries, ranging from 3 to 9, reflecting the diverse scope of the selected topics.

Table S15: Mode B per-query results for the Infectious Diseases domain (hybrid retrieval with cross-encoder re-ranking). $|\mathcal{R}|$ denotes the number of manually annotated relevant chunks.

Query ID	$\|\mathcal{R}\|$	MRR	R@1	R@5	R@10	nDCG@5
Q1	3	1.000	0.333	0.667	0.667	0.765
Q2	5	0.250	0.000	0.200	0.200	0.146
Q3	9	1.000	0.111	0.333	0.333	0.699
Q4	6	1.000	0.167	0.167	0.167	0.339
Q5	4	0.000	0.000	0.000	0.000	0.000

The queries cited in Table S15 are listed below:

Q1: *What is the role of antimicrobial resistance in infection prevention?*

Q2: *How can precision-medicine approaches help predict stress-granule–linked antimicrobial resistance mechanisms?*

Q3: *How does early immune imbalance drive the shift from inflammation to immunosuppression in septic patients?*

Q4: *Which are the consequences of global changes in disease transmission?*

Q5: *What molecular mechanisms regulate the stability and immune function of the tuberculosis granuloma?*

Query 5 (Q5) achieves zero retrieval across all cutoffs despite having 4 relevant chunks. This query targets the molecular mechanisms of granuloma stability, a topic whose relevant content is distributed across multiple documents covering different aspects of tuberculosis immunology, making it difficult for the retriever to surface any single relevant chunk. Query 3 (Q3), despite having the largest relevant set ($|\mathcal{R}|$=9), achieves MRR = 1.0 but low recall at all cutoffs (R@5 = 0.333), as only 3 of the 9 relevant chunks are retrieved within the top-10 positions. Query 2 (Q2) also struggles, retrieving only 1 of 5 relevant chunks at rank 4, reflecting the specialized nature of the query about stress-granule–linked AMR mechanisms. The aggregate results across all domains are summarised in the Mode B aggregate-results table of the main manuscript.

S12 Interactive Q&A

Table S16 summarises the representative interactive Q&A sessions for the three biomedical domains not discussed in the main text (Diabetes, Pharmacology, and Infectious Diseases). For each domain, the query achieving the highest mean reranker score ($\bar{s}$) was selected as a representative example. All selected cases were assessed as *Correct*, indicating that highly ranked supporting passages consistently led to accurate grounded answers.

For the Diabetes domain, Query 3 (*What mechanisms link metabolic dysfunction to inflammation in type 2 diabetes?*) achieved the highest mean reranker score ($\bar{s} = 0.959$), with all three retrieved passages scoring above 0.88.

For the Pharmacology domain, Query 4 (*What is pharmacogenomics and how can it help personalize drug therapy?*) obtained a mean reranker score of $\bar{s} = 0.982$, with all retrieved passages scoring above 0.96.

For the Infectious Diseases domain, Query 4 (*What global changes in the past two decades have most influenced infectious disease dynamics?*) achieved the highest overall confidence among the three domains ($\bar{s} = 0.987$), with all retrieved passages scoring above 0.98.

Table S16: Interactive Q&A summary for the Diabetes, Pharmacology, and Infectious Diseases domains (hybrid retrieval with cross-encoder re-ranking). $\bar{s}$: mean reranker score; Quality: C = Correct, P = Partial, F = Failed. The Oncology Q&A results are reported in the Oncology interactive Q&A table in the main manuscript of the main text.

Domain	**Q**	**Query**	**#Src**	$\bar{s}$	**Quality**
Diabetes	Q3	What mechanisms link metabolic dysfunction to inflammation in T2D?	3	0.959	**C**
Pharmacology	Q4	What is pharmacogenomics and how can it help personalize drug therapy?	3	0.982	**C**
Infectious Diseases	Q4	What global changes in the past two decades have most influenced infectious disease dynamics?	3	0.987	**C**

References

[1] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759, 2009.

[2] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[3] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.