Review Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2026.66.010413

Small Language Models for Chronic Diseases: Managing Diabetic Retinopathy in the Global South

Ravi S Sharma*, Edmund Evangelista, Sheik Mohammed Ali, Salman Bukhari

College of Technological Innovation, Zayed University, Abu Dhabi & Dubai, UAE

***Corresponding author:** Ravi S Sharma, Center for Inclusive Digital Enterprise (CeIDE), Wellington, New Zealand and College of Technological Innovation, Zayed University, Abu Dhabi & Dubai, UAE

ARTICLE INFO

Received: September 18, 2026

Published: September 29, 2026

Citation: Ravi S Sharma, Edmund Evangelista, Sheik Mohammed Ali, Salman Bukhari. Small Language Models for Chronic Diseases: Managing Diabetic Retinopathy in the Global South. Biomed J Sci & Tech Res 66(5)-2026. BJSTR. MS.ID.010413.

ABSTRACT

Access to reliable chronic disease knowledge remains uneven in settings with limited specialist availability, computing resources, connectivity, and funding for developing artificial intelligence (AI) systems. This Paper presents a frugal and responsible knowledge-support framework using diabetic retinopathy (DR) as a use-case domain. The framework converts a curated medical research collection into cleaned passages, clinical entities, knowledge graph (KG) relationships, and compact semantic embeddings stored in a local SQLite database. User questions first pass through scope and safety screening before semantic, keyword, and KG retrieval assemble relevant evidence. Responses can then be produced through a local template or a small language model (SLM), with sources and an education-only boundary retained. The shared knowledge base of our Proof-of-Concept (PoC) contained 160 papers, 6,431 embedded passages, 16,484 entity records, and 6,080 relationships. The same architecture was adapted to two complementary scenarios (prevention counselling and treatment evidence access). These adaptations were achieved mainly by changing the medical research corpus, terminology, relationships, and question sets rather than by training separate large models. The PoC demonstrates the feasibility of building an inspectable, locally stored, and optional clinical support system using modest infrastructure. It does not provide diagnosis, treatment selection, or evidence of clinical effectiveness. Further work would strengthen evaluation, multilingual support, urgent-care routing, evidence updating, and local generative small-model integration.

Keywords: Frugal Artificial Intelligence; Responsible Artificial Intelligence; Diabetic Retinopathy; Chronic Disease Clinical Support; Retrieval-Augmented Generation; Knowledge Graphs; Small-Model Retrieval; Global South

Abbreviations: AI: Artificial Intelligence; KG: Knowledge Graph; SLM: Small Language Model; PoC: Proof-Of-Concept; DR: Diabetic Retinopathy; SR: Systematic Review; WHO: World Health Organization; GAI: Generative Artificial Intelligence; RAG: Retrieval Augmented Generation; GS: Global South; LLMs: Language Models

Introduction

Diabetes is a rapidly growing global health challenge. The International Diabetes Federation estimated that 589 million adults were living with diabetes in 2024 and projected this number to reach 853 million by 2050 [1]. As more people live longer with diabetes, the burden of related complications will also increase. Diabetic retinopathy (DR) is among the most critical of these complications because it can progress without early symptoms and may lead to avoidable visual impairment when screening and treatment are delayed [2]. A global systematic review (SR) estimated that DR affected 22.27% of people with diabetes, representing approximately 103 million adults in 2020. This number was projected to rise to more than 160 million by 2045. The burden is not evenly distributed, with particularly high prevalence and projected growth in several regions that include low-income and middle-income countries [3]. Prevention, regular retinal assessment, early detection, and timely treatment are therefore important parts of the wider diabetes care direction [2,4]. Despite this need, access to DR knowledge and services remains uneven. The World Health Organization (WHO) has emphasized integrating evidence-based eye care interventions into health systems, particularly in low-income and middle-income countries [4].

However, national planning for DR screening remains limited in many such settings, with reported challenges including weak service integration, limited long-term funding, unclear referral pathways, shortages of trained personnel, and gaps in public and professional awareness [5,6]. These challenges affect not only screening and treatment but also access to clear and reliable knowledge for educators, students, primary care teams, and people living with diabetes. Recent advancements in generative artificial intelligence (GAI) offer new ways to search, summarize, and communicate medical information [7]. However, unrestricted reliance on large general purpose language models (LLMs) is difficult to justify in health contexts. Their responses may be difficult to trace to specific evidence, and their deployment may depend on substantial computing resources, commercial services, reliable connectivity, and external data processing [8,9]. The WHO has therefore emphasized that AI for health should be designed around transparency, accountability, safety, human oversight, and protection of individual and public interests [8,10]. Retrieval augmented generation (RAG) offers a more grounded alternative by connecting language generation to external evidence collection rather than relying only on knowledge stored within model parameters [11]. Compact transformer models can also reduce the computational demands of semantic text processing while retaining useful language representations [12,13].

A related diabetes focused PoC has combined peer-reviewed literature, KG construction, graph-based retrieval, and a local language model to support evidence-grounded knowledge access [14]. These developments create an opportunity to place evidence retrieval, structured knowledge, and safety controls before generation. However, their responsible use in health requires clear boundaries: retrieved evidence may still be incomplete, knowledge graph (KG) links may not imply causality, and fluent output must not be treated as personal diagnosis or treatment advice [8,10]. The proposed approach, as illustrated in Figure 1, responds to challenges in evidence access, specialist availability, connectivity, computing resources, model cost, and answer reliability through an affordable, grounded, inspectable, adaptable, and safety-bounded design. Against this background, this Paper presents a frugal and responsible knowledge support framework using DR as a PoC domain. The framework converts a bounded collection of literature into cleaned passages, clinical entities, KG relationships, and compact semantic embeddings stored in a local SQLite database. At query time, safety screening is applied before hybrid passage and graph retrieval. Retrieved evidence is then presented through a local template or an optional language model, with sources and an education-only boundary.

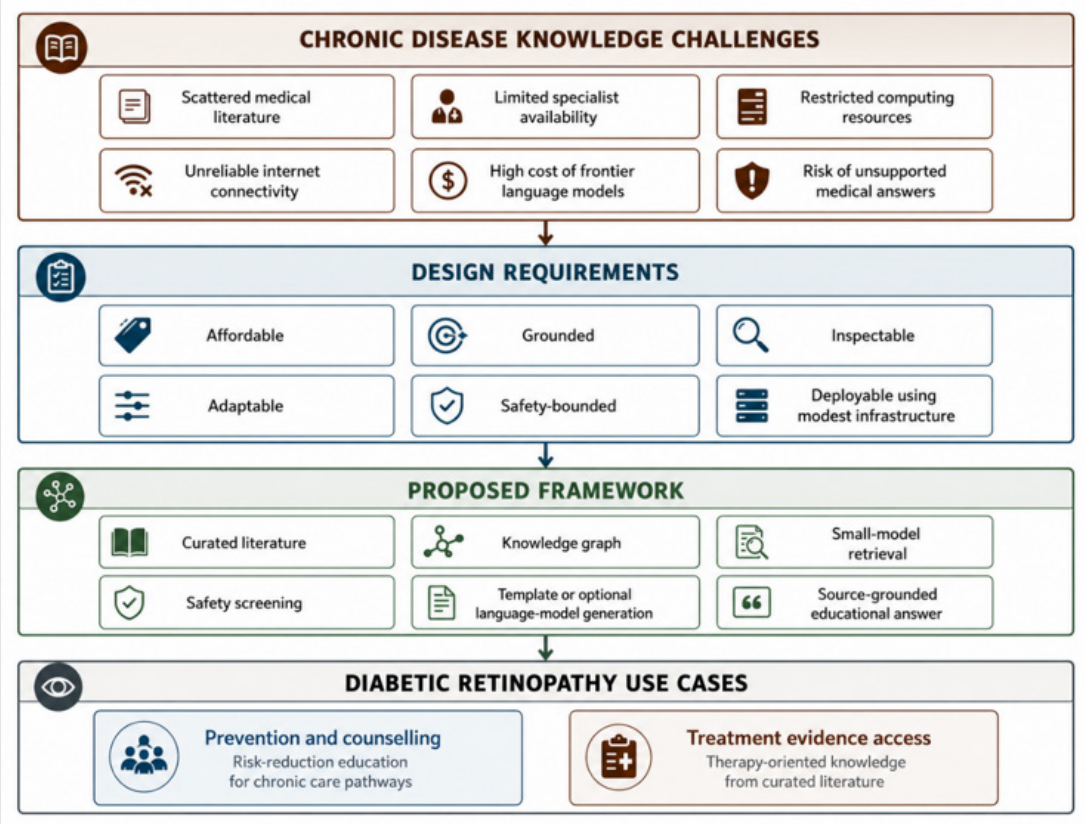

Figure 1: From chronic disease challenges to a frugal and responsible AI framework.

The Paper uses one shared implementation as its technical core and examines two adaptations (prevention counselling and treatment evidence access). These adaptations reuse the same architecture while changing the corpus, terminology, relationship emphasis, and question set. The purpose is not to develop a diagnostic system or replace clinical services. Rather, the work explores whether curated evidence, structured knowledge, lightweight retrieval, and explicit safety controls can provide a practical foundation for chronic disease education in resource-constrained settings. The contribution is therefore threefold. First, the Paper provides an operational account of frugal and responsible health knowledge support rather than treating these terms as general aspirations. Second, it shows an end-to-end DR implementation using open tools, local storage, compact retrieval, and optional generation. Third, it shows how the same framework can be adapted to prevention-based and treatment-based knowledge needs without making repeated large model training the default approach. The framework is described as Global South (GS)-based, not as a clinically validated deployment across GS. Its relevance comes from its emphasis on modest infrastructure, local inspectability, open tools, intermittent connectivity potential, and capacity building [9,15,16]. Its clinical usefulness, safety, multilingual accessibility, and real-world acceptability will require stronger evaluation before any patient facing application [8,10].

The remainder of this Paper is organized as follows. Section 2 reviews related work on RAG, medical KGs, compact language models, and responsible health AI. Section 3 presents the frugal and responsible design goals. Section 4 describes the framework and its DR PoC implementation. Section 5 explains the prevention and treatment adaptations, including the worked retrieval example. Section 6 discusses the main findings, limitations, future priorities, and concluding remarks.

Background and Related Work

The technical foundation of this Paper lies at the intersection of RAG, biomedical KGs, compact language models, and responsible AI. These areas offer complementary strengths. Retrieval provides access to external evidence, KGs organize clinical concepts, compact models reduce computational demand, and responsible AI principles define how the system should behave in a health context.

Retrieval Augmented Health Information

RAG was introduced to connect language generation with external knowledge rather than relying only on information stored within model parameters. A typical RAG system indexes a document collection, retrieves passages relevant to a user's question, and provides those passages to a language model during answer construction [11]. This approach can make evidence updates easier because new documents can be added to the retrieval collection without retraining the complete model. Health related studies have increasingly applied this approach to clinical guidelines and patient education [17], for example, showed that restructuring hepatology guidelines and providing them as retrieved context substantially improved the accuracy of language model responses. In diabetes education, [12] evaluated a RAG system across 43 common questions and reported improvements in response accuracy, completeness, and understandability across several general-purpose language models. Similarly, the Almanac framework connected language model responses to curated medical resources, showing how retrieval can support source grounded answers to clinical questions [18].

These findings support the value of retrieval for domain specific health questions. However, RAG does not make a health system automatically reliable. Its output still depends on the quality of the source collection, document processing, retrieval method, ranking strategy, and generation prompt [19]. Incorrect or weakly related passages can still lead to misleading answers. Responsible use therefore requires source traceability, clear scope boundaries, human review, and safety controls in addition to retrieval itself [8,10]. These requirements are consistent with the WHO's emphasis on transparency, accountability, safety, and human oversight in health AI.

Knowledge Graphs for Structured Medical Evidence

Biomedical KGs represent information as entities and relationships [20]. In a DR context, entities may include diseases, risk factors, disease stages, treatments, drugs, biomarkers, and outcomes. Relationships can then connect concepts such as a risk factor with a complication or a treatment with disease manifestation. KGs can complement passage retrieval by making important concepts easier to locate and inspect. [21] showed this approach by constructing a diabetes complications KG from expert-reviewed clinical evidence. Their work showed how disease, risk factor, and outcome information can be transformed into a structured representation rather than retained only as unconnected text. More recently, [14] combined a locally operated language model with KG-supported retrieval in a gestational diabetes PoC, showing how graph structures and retrieved literature can be integrated within an evidence-grounded question-answering workflow. In this Paper, the value of a KG is mainly organizational. It helps users and developers understand which concepts were identified and how they were connected in the retained literature. It also supports adaptation because new terms and relation types can be added without retraining the complete retrieval system.

However, Graph relationships must be interpreted carefully. A relationship produced through co-occurrence or a matching rule does not necessarily establish causality, treatment effectiveness, or clinical recommendation. The original passage and source remain necessary for interpretation. The framework therefore uses the graph to guide retrieval and improve transparency, not as an autonomous clinical reasoning engine.

Compact Models and Frugal Language Processing

Large transformer models can require substantial memory, com-

putation, and serving infrastructure [22]. Model compression methods provide one route to reducing these requirements. MiniLM uses knowledge distillation to create smaller transformer representations while retaining much of the performance of larger teacher models across language understanding tasks [12]. The design framework uses all-MiniLM-L6-v2 as a compact sentence embedding model. Its role is limited to representing questions and passages for semantic retrieval. It is not used for personal diagnosis, treatment selection, or independent clinical reasoning [12,13]. This limited use is important to the Paper's definition of small model AI. The system does not claim that every answer is generated by a locally deployed small generative model. Instead, it combines compact retrieval with local storage and template-based answer construction. An external language model can be added for smoother wording, but it is not required for evidence retrieval or source presentation. This architecture differs from many medical RAG studies that retain a large cloud-based language model as the central answering component [17-19].

Previous work has shown that RAG can improve the performance of such models, but fewer implementations treat generation as optional while combining local storage, controlled terminology, KG retrieval, and pre-answer safety screening. This gap is especially relevant where continuous connectivity, commercial APIs, and high-performance infrastructure cannot be assumed [9,15]. This comparison is an inference from the selected RAG and KG studies rather than a claim from a formal systematic review.

Positioning of the Design Framework

Existing work establishes the value of external retrieval, structured clinical knowledge, and compact language representations [11,13,20]. The framework presented in this Paper brings these elements together around four priorities:

- Bounded and locally stored literature collection.
- Hybrid passage and KG retrieval.
- Template first and optional language model answering.
- Safety screening before a medical response is produced.

Its contribution is therefore not a new foundational model or a clinical decision system. It is an integrated PoC architecture showing how existing methods can be combined into a more inspectable and resource aware form of chronic disease knowledge support. The framework is also adapted through changes to evidence and terminology rather than repeated large model training. This is shown through its prevention counselling and treatment evidence applications. The next section defines the frugal and responsible principles that guided these implementation choices.

Design Goals: Frugal and Responsible

The framework was designed around a simple idea: useful health knowledge should not depend on expensive computing infrastructure, continuous internet access, or unrestricted use of LLMs. This is especially important in resource constrained settings, where specialist support, technical capacity, and funding may be limited. In such environments, an AI system must do more than produce fluent text. It should also be affordable, maintainable, transparent, evidence grounded, and clear about what it can and cannot do [8,10,15]. Frugality and responsibility were therefore treated as connected design goals. A low-cost system is not useful if it produces unreliable medical information, while a responsible system is difficult to sustain if it depends on infrastructure that intended users cannot access. The PoC combines lightweight technologies with source grounding, inspectable knowledge structures, safety checks, and an explicit 'education not diagnosis' boundary. Table 1 presents how these principles were implemented [8,11,12,21].

Table 1: Operationalization of frugal and responsible design principles.

Design Principle	Operationalization in the Proof of Concept
Frugal computation	Open Python tools and a compact embedding model that can operate on ordinary computing hardware
Frugal storage	Portable SQLite databases requiring no separate database server
Limited dependence on large models	Template first answering with optional language model wording
Evidence before generation	Answers are based on retrieved passages from a bounded literature corpus
Transparency	Editable dictionaries, inspectable relationships, retrievable passages, and local query outputs
Safety before answering	Personal diagnosis and treatment selection requests are screened before generation
Knowledge-based adaptation	Prevention and treatment scenarios are specialized through corpus and terminology changes
Global South orientation	Local storage, modest infrastructure, open tools, and opportunities for local capacity building

Frugal System Design

Generally speaking, frugal systems are designed to fulfil the satisficing principle rather than provide optimized solutions. The first goal was to reduce unnecessary dependence on high-performance computing [15,22]. The framework uses open-source tools and does not require the training or fine-tuning of a large generative model. Semantic retrieval is supported by all-MiniLM-L6-v2, a compact sentence embedding model that represents user questions and literature passages in a shared semantic space [12,13]. Its role is limited to iden-

tifying relevant evidence. It is not used to diagnose disease, predict outcomes, or independently recommend treatment. The processed papers, passages, embeddings, entities, and relationships are stored in SQLite. This lightweight database does not require a separate server and can be retained as a single local file. The shared DR database occupies approximately 64.4 MB, while the treatment and prevention adaptations occupy 23.6 MB and 16.5 MB, respectively. These modest sizes support local storage, backup, inspection, and transfer between ordinary computing environments. The system also uses a template-first answer pathway. Retrieved evidence can be organized into an educational response without sending the question or source material to an external service.

This keeps the main workflow usable when internet access, API funding, or cloud services are unavailable. An external language model can be added to improve readability, but it is treated as an optional wording layer rather than the main source of knowledge. This design follows a 'small before large' philosophy. The system first identifies concepts, retrieves passages, examines KG relationships, and assembles the supporting evidence. Language generation is introduced only after these steps. The PoC therefore shows small model retrieval and structured evidence processing rather than claiming that every answer was produced by a locally deployed generative small language model. Because the architecture is modular, a local generative model could later replace the external service without rebuilding the corpus, graph, embeddings, or safety layer.

Evidence, Transparency and Safety

The main responsible design principle is that answers should begin with evidence rather than model memory [11,18]. The framework searches a bounded DR corpus and retrieves relevant passages before building a response. This creates a traceable path from the question of the user to the literature used in the answer. Retrieval does not guarantee that every source is complete or correct, but it makes the evidence visible and reviewable [8,19]. This is particularly important in healthcare, where a fluent response may appear trustworthy even when it is unsupported. Separating retrieval from generation lets users and developers inspect whether the response connects to relevant literature, rather than judging it only by tone or confidence. The framework adds a KG to organize clinically meaningful concepts, including diseases, disease stages, risk factors, treatments, drugs, biomarkers, and treatment outcomes [20,21]. These concepts are identified using controlled dictionaries and matching rules. The terminology and relationships can be reviewed and edited directly, making the system more transparent than one that depends entirely on opaque model behaviour. The graph is used for organization and retrieval, not as a complete clinical ontology.

Relationships based on co-occurrence or predefined rules should not be interpreted as proof of causality or treatment effectiveness. Their purpose is to show how concepts are connected within the retained literature and to help the system locate relevant evidence. Safety controls are applied before generating an answer. The system is designed for education and knowledge support, not for personal diagnosis, medication changes, or treatment selection. It may explain the general role of anti-vascular endothelial growth factor therapy, for example, but it should not decide whether a particular person should receive a specific injection. Questions requesting personal diagnosis, treatment selection, medication changes, or unsupported health predictions are refused or redirected to qualified care. A disclaimer is also added to the output, but it is treated as a secondary protection. The main protection is blocking an inappropriate request before producing a detailed answer. Urgent symptoms require an even stronger boundary. A question describing sudden vision loss should prioritize immediate professional care rather than return a long literature summary. The current system recognizes this as a design requirement, although urgent care routing still requires broader testing before real-world use.

Transparency also supports system maintenance. When retrieval is weak, the team can examine the corpus, cleaned text, dictionary, graph relationships, or ranking logic. In the treatment adaptation, we addressed weak domain coverage by expanding the dictionary from approximately 84 to 190 terms and adding categories for drugs, biomarkers, and treatment outcomes. This measure-revise-reassess process is more inspectable than attempting to solve every weakness through larger-model prompting.

Adaptation and Global South Orientation

A further goal was to make the framework reusable across related chronic care scenarios. The prevention and treatment adaptations retained the same core architecture but differed in their evidence collections, terminology, relationships, and questions. The prevention adaptation emphasized glycaemic control, blood pressure, lipid management, pregnancy, screening, and modifiable risk factors. The treatment adaptation emphasized anti-vascular endothelial growth factor therapy, laser procedures, vitrectomy, drugs, biomarkers, and treatment outcomes. These differences were introduced mainly by adapting the corpus and structured knowledge rather than training separate large models. This approach is relevant to resource-constrained environments of Global South because repeated model training can be expensive, technically demanding, and difficult to maintain. Corpus and dictionary updates are comparatively easier to review, document, and improve over time. They also allow domain experts and junior researchers to contribute directly to system development. The framework is described as satisficing rather than as clinically validated optimal deployment.

Its use of open tools, local storage, CPU-compatible retrieval, optional cloud generation, and modular adaptation respond to common limitations such as limited infrastructure, variable connectivity, restricted specialist access, and the need for locally maintainable systems. The development process also supported capacity building. Junior researchers could participate in literature curation, terminology development, knowledge base construction, and question testing without needing to train a complex language model. This shifts AI

work from passive use of commercial systems toward a clearer understanding of how evidence, retrieval, and safety interact. Important boundaries remain. The current literature is mainly English-language and influenced by open-access availability. The framework has not been clinically deployed, evaluated across health systems, or shown to improve patient outcomes. It should therefore be understood as an educational and technical PoC with potential relevance to resource-constrained settings. The design places curated evidence, structured knowledge, lightweight retrieval, and safety controls before generative complexity. This position guides the end-to-end framework described in the next section.

Framework and Proof-of-Concept Implementation

The design framework converts a bounded collection of DR literature into a searchable and inspectable knowledge support system. It separates knowledge preparation from real-time question answering. In the offline stage, selected publications are cleaned, divided into passages, enriched with clinical concepts and relationships, embedded using a compact semantic model, and stored locally. In the online stage, a user's question first passes through safety and scope checks. The system then retrieves relevant passages and KG links before presenting the evidence through a local template or an optional language model. Figure 2 provides an overview of these two connected stages. The offline knowledge refinery converts selected literature into cleaned passages, clinical entities, relationships, semantic embeddings, and a local SQLite knowledge base. The online workflow applies safety screening, hybrid retrieval, KG retrieval, evidence assembly, and template first or optional language model wording. The generative component is therefore an optional presentation layer rather than the complete system.

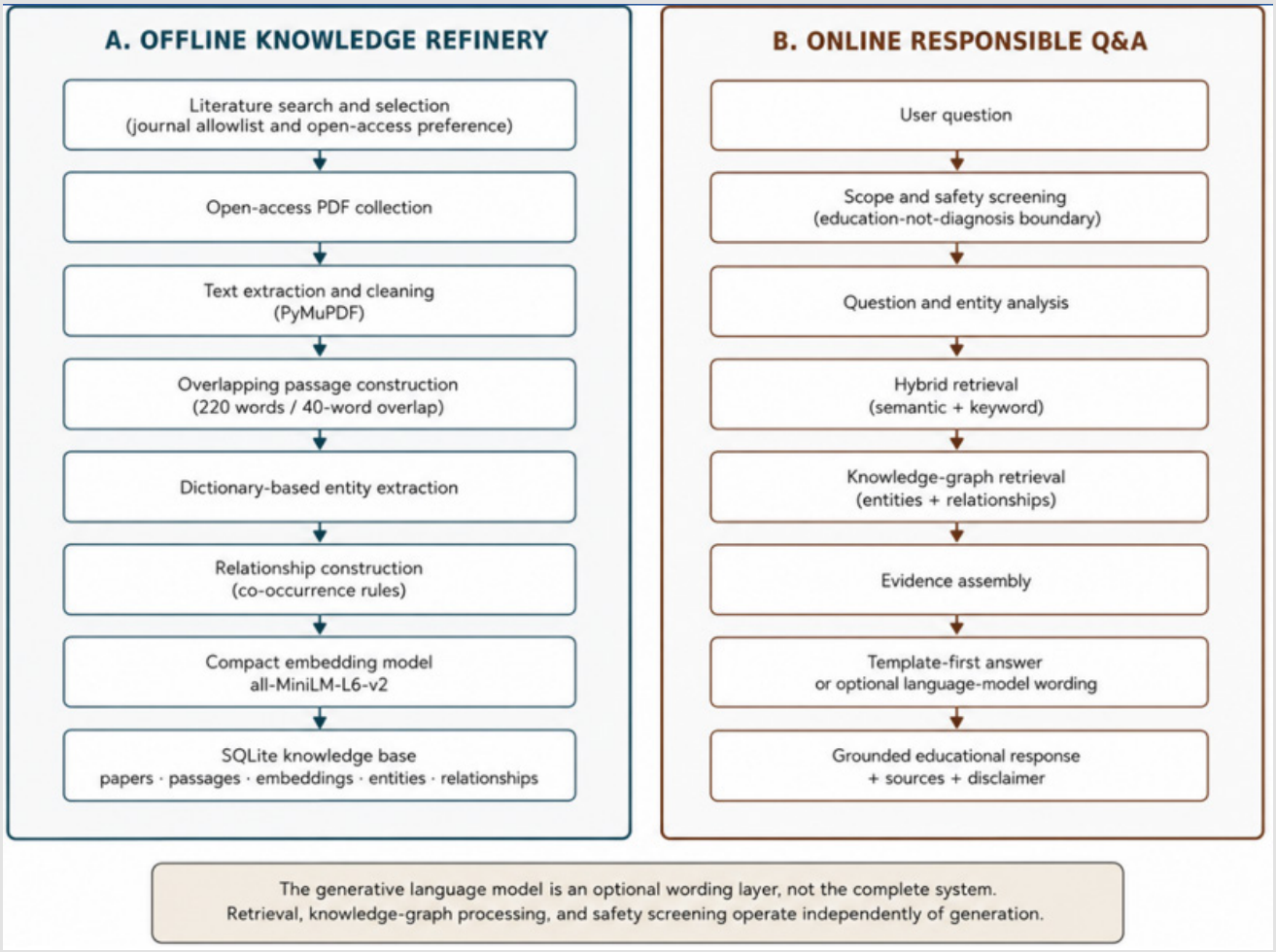

Figure 2: End-to-end architecture of the frugal and responsible DR clinical assistant.

Offline Knowledge Refinery

The offline stage prepares the evidence base before any user question is submitted. This reduces query-time processing and allows the underlying literature, terminology, and knowledge structures to be reviewed and updated independently. The process begins with literature collection. Relevant DR studies were identified through Semantic Scholar searches, a journal allowlist, and manual relevance review. The project prioritized open-access papers because they could be downloaded, processed, and retained locally without continuous access to commercial publisher platforms. The final shared database contained 160 papers published between 2020 and 2024, covering risk factors, screening, disease stages, diabetic macular edema, preventive management, and treatment approaches. The use of a bounded corpus was deliberate. The framework does not search the entire internet whenever a question is asked. Instead, it retrieves from a defined collection that has already passed through the project's selection and processing workflow. This makes the evidence environment easier to inspect and improves traceability. However, it also means that the quality and coverage of the system depend on the selected literature.

Open-access preference and journal-based filtering may exclude relevant subscription-based studies, clinical guidelines, regional publications, and implementation evidence. The collection should therefore be understood as a proof-of-concept corpus rather than a complete systematic review. Selected PDFs were converted into machine-readable text using PyMuPDF. Cleaning reduced repeated headers, footers, spacing problems, page artefacts, and other noise introduced during PDF extraction. The cleaned text files were retained as intermediate outputs so that obvious extraction errors could be identified before indexing. Each document was then divided into passages of approximately 220 words, with an overlap of 40 words between adjacent passages. The overlap helps preserve statements that cross a passage boundary, while the selected length provides enough context for interpretation without requiring the system to retrieve a complete paper. These passages became the main units for semantic search and answer grounding. The next stage added a structured knowledge layer. Controlled dictionaries identified concepts such as diseases, disease stages, risk factors, symptoms, screening methods, treatments, drugs, biomarkers, and treatment outcomes. Dictionary-based extraction was selected because it is lightweight and easy to inspect.

Terms can be added, corrected, or removed without retraining a clinical information-extraction model. Relationships were constructed when relevant concepts appeared together or satisfied predefined rules. For example, a passage discussing hypertension in relation to DR could support a risk related link, while a passage connecting anti-vascular endothelial growth factor therapy with diabetic macular edema could support a treatment related link. These relationships help organize the literature and support concept-based retrieval. The resulting KG is not presented as a complete clinical ontology. Many links reflect textual co-occurrence or rule-based associations and should not be interpreted as proof of causality, treatment effectiveness, or universal clinical guidance. Its purpose is practical (to make the evidence easier to organize, retrieve, and inspect). Each passage was also converted into semantic embedding using all-MiniLM-L6-v2. The same compact model represents user questions, allowing the system to retrieve conceptually related passages even when the wording differs. For example, a question about "preventing eye complications from diabetes" may still retrieve literature discussing "reduction of DR progression."

The embedding model is used only for retrieval. It does not diagnose disease, predict outcomes, or make treatment decisions. Passage embeddings are generated once during database construction and stored locally, so only the incoming question must be embedded during an online query. The papers, passages, embeddings, entities, relationships, and source links are stored in SQLite. This server-free format allows each knowledge base to be retained as a single local file, supporting backup, inspection, transfer, and possible use under limited-connectivity conditions.

Online Responsible Question Answering

The online stage begins when a user submits a question. Before retrieving evidence, the framework checks whether the request falls within its educational scope. General questions about DR, risk reduction, screening, or treatment evidence are allowed to proceed. Requests for personal diagnosis, individual treatment selection, medication changes, or unsupported predictions are refused or redirected. The system may explain how anti-vascular endothelial growth factor therapy is generally used, but it should not decide whether a particular person should receive a specific injection. This safety-first order is important. Responsibility should not depend only on adding a disclaimer after an unrestricted answer has already been produced. The primary protection is the decision made before retrieval and generation about whether the request is appropriate. The current safety layer uses pattern-based and scope-based rules. These rules can be inspected and revised, but they do not guarantee complete coverage. Users may express unsafe requests indirectly or in unexpected language. Questions describing sudden vision loss or severe visual change also require stronger urgent care routing. Such questions should prioritize immediate professional assessment rather than return a long literature summary. This behaviour remains a design requirement that needs broader testing before real-world deployment. Once a question is permitted, the system retrieves evidence through three complementary routes:

1. **Semantic Retrieval:** Which identifies passages with similar meaning.
2. **Keyword Retrieval:** Which preserves exact clinical terms, drug names, and procedure names.
3. **Kg Retrieval:** Which identifies relevant entities and relationships.

These approaches address different limitations. Semantic retrieval helps when the question and literature use different wording. Keyword matching is important for exact terms such as drug names or abbreviations. KG retrieval adds structured context by linking the concepts found in the question to related risks, treatments, biomarkers, or outcomes. The retrieved results are ranked and assembled into a compact evidence package containing the most relevant passages, detected concepts, graph relationships, source identifiers, and available relevance information. Only this selected evidence is passed to the answering component, rather than the full literature collection. The default answer pathway uses a local template. A template-based response can contain a concise educational explanation, important clinical concepts, supporting relationships, retrieved passages, source details, and a scope disclaimer. Because it reorganizes retrieved material rather than freely generating medical claims, its output is relatively predictable and does not require an external generative service. An optional language model can be used to improve readability and combine the retrieved evidence into smoother prose.

However, the language model receives the evidence only after retrieval and is expected to remain within that evidence boundary. The system is therefore retrieval-first rather than model-first. The corpus, KG, safety checks, and template mode remain usable when the external generator is unavailable or replaced. The final output is evidence grounded educational response rather than a clinical decision. Depending on the available metadata, it may include paper titles, article identifiers, passage references, or other traceable source information. It also explains that the response does not replace screening, diagnosis, specialist assessment, treatment planning, or professional judgement.

Quantitative Implementation Footprint

The shared implementation served as the canonical PoC system. Two smaller knowledge bases were later created to examine whether the same architecture could support treatment evidence access and prevention counselling. Table 2 presents the verified footprint of all three databases. All passages in the three databases had stored semantic embeddings. As shown in Table 2, the shared system remained small enough for local storage despite containing more than 6,000 retrievable passages and over 16,000 entity records. The treatment and prevention databases were smaller adaptations rather than competing primary systems. Their role was to show how the knowledge layer could be specialized while preserving the same core architecture. The treatment adaptation also shows the value of controlled terminology. Its dictionary was expanded from approximately 84 to 190 terms and strengthened with categories for drugs, biomarkers, and treatment outcomes. This improved the representation of treatment focused literature without changing the embedding model, storage layer, or overall retrieval workflow.

Table 2: Knowledge-base footprint of the shared framework and its two adaptations.

Knowledge Base	Clinical Focus	Papers	Passages	Extracted Entity Records	Relationships	Embedded Passages	Database Size
Shared main knowledge base	General diabetic-retinopathy knowledge	160	6,431	16,484	6,080	6,431	64.4 MB
Treatment adaptation	Treatment and therapeutic evidence	62	2,346	6,388	2,562	2,346	23.6 MB
Prevention adaptation	Prevention and counselling knowledge	39	1,640	2,986	787	1,640	16.5 MB

Implementation of Components and Their Roles

Table 3 presents the main technical components and shows how each contributes to the frugal and responsible design of the framework. As Table 3 shows, no single model is treated as the system's complete intelligence. The implementation relies on a sequence of bounded operations, beginning with evidence selection and ending with evidence-based presentation.

Table 3: Components of the PoC framework and their implementation roles.

Framework Component	Tool or Method	Main Function	Frugality Contribution	Responsibility Contribution
Literature collection	Semantic Scholar search, journal allowlist, and manual review	Build a bounded evidence corpus	Prioritizes accessible literature and reusable metadata	Makes the evidence base identifiable and reviewable
PDF processing	PyMuPDF-based extraction and cleaning	Converts articles into usable text	Uses an open local processing library	Allows extracted content to be inspected before indexing
Passage construction	Approximately 220 words with 40-word overlap	Creates focused retrieval units	Uses simple and low-cost preprocessing	Supports passage level evidence grounding
Entity extraction	Controlled dictionaries and matching rules	Identifies diseases, risks, treatments, drugs, biomarkers, and outcomes	Avoids training a large extraction model	Terminology can be reviewed and edited
Relationship construction	Rule-based and co-occurrence logic	Connect relevant clinical concepts	Provides lightweight graph construction	Relationships remain inspectable and exportable
Semantic embeddings	all-MiniLM-L6-v2 [12]	Represent passage meaning for retrieval	Compact and CPU compatible	Ensures retrieval occurs before generation
Local storage	SQLite	Store papers, passages, embeddings, entities, and relationships	Portable and server free	Supports local custody and reproducibility
Hybrid retrieval	Semantic, keyword, and knowledge-graph search [11]	Assemble question-relevant evidence	Avoids processing the full corpus for each query	Reduces reliance on unsupported model memory
Safety screening	Pattern-based and scope-based rules	Detect inappropriate personal medical requests	Adds little computational overhead	Blocks requests outside the educational boundary
Answer construction	Local template or optional language model	Present retrieved evidence clearly	Remains usable without a generative API	Retains sources and disclaimers in the output

Modular and Reproducible Workflow

The framework was implemented as a staged pipeline rather than a single monolithic application. The workflow covered literature collection, PDF cleaning, passage construction, entity extraction, relationship development, embedding generation, database construction, hybrid retrieval, safety checks, and answer presentation. Each stage produced outputs that could be inspected before running the next stage. This made it easier to identify whether weak retrieval resulted from the source documents, PDF extraction, passage boundaries, missing terminology, graph relationships, or ranking logic. The shared configuration files retained settings such as passage length, overlap, embedding model, and database locations. The controlled dictionaries were stored separately so that terminology could be expanded without modifying the full codebase. The framework therefore contains both shared and adaptable elements. The shared elements include text processing, passage construction, embeddings, SQLite storage, hybrid retrieval, safety screening, and answer assembly. The adaptable elements include the literature corpus, controlled terminology, entity categories, relationship rules, and intended question set. For responsible dissemination, exclude environment files, API credentials, and other secrets from any public archive. Reproducibility materials should instead include dependency requirements, configuration examples, database-building instructions, and sample questions.

Intended Use and Boundaries

The framework is intended to support literature navigation, professional and student education, supervised patient education development, prevention counselling preparation, treatment evidence review, and responsible artificial intelligence training. It is not intended to replace retinal screening, specialist examination, clinical guidelines, diagnosis, prescribing, or professional judgement. Its contribution is showing how a bounded medical evidence collection can be transformed into an accessible, inspectable question-answering resource using modest technical infrastructure. The next section examines how this shared framework was adapted to prevention counselling and treatment evidence access through changes to the corpus, terminology, relationships, and question sets.

Scenario-Based Adaptation of the Shared Framework

The PoC was adapted to two related knowledge needs (treatment evidence access and prevention counselling). These were not separate AI systems. Both reused the same workflow for literature processing, passage construction, KG development, MiniLM-based retrieval, SQLite storage, safety screening, and grounded answer construction. The main changes concerned the literature included, the terminology represented, the graph relationships emphasized, and the questions asked. This approach supports the Paper's central argument. The specialization can be achieved by adapting evidence and

domain knowledge rather than repeatedly training large models. Figure 3 shows how the shared framework was extended to the two DR scenarios. The two adaptations retain the same technical core but differ in their corpora, terminology, graph emphasis, and question sets.

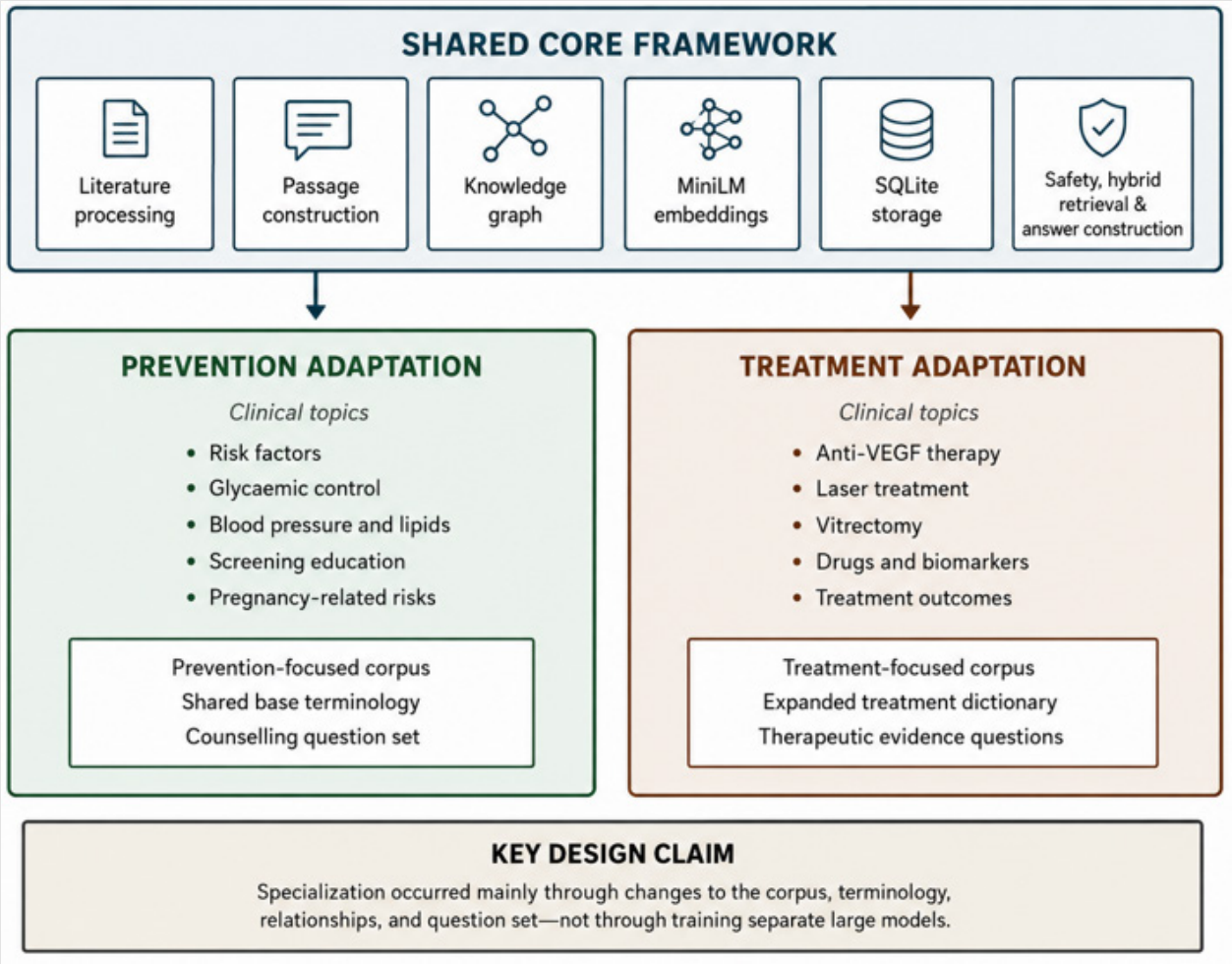

Figure 3: Adaptation of framework for prevention counselling and treatment evidence access.

Treatment Evidence Adaptation

The treatment focused adaptation examined whether the shared framework could support educational questions about DR therapies. Its literature searches emphasized anti-vascular endothelial growth factor therapy, laser procedures, vitrectomy, drugs, biomarkers, and treatment outcomes. After manual relevance review, the treatment database contained 62 papers and 2,346 passages, producing 6,388 entity records and 2,562 relationships. All passages had stored semantic embeddings, and the complete SQLite database occupied approximately 23.6 MB. The original shared dictionary covered general DR concepts but lacked sufficient detail for specialized treatment questions. It was therefore expanded from approximately 84 to 190 terms, with stronger coverage of drugs, biomarkers, procedures, and treatment outcomes. This change improved the structured representation of the treatment literature without retraining the embedding model or changing the database architecture. The adaptation was designed for questions such as:

- What is anti-VEGF therapy used for?
- How is diabetic macular edema commonly treated?
- What role does laser treatment play in diabetic retinopathy?
- When is vitrectomy discussed in the literature?
- Which biomarkers are studied in relation to treatment response?

These are educational questions rather than personal treatment requests. The system may summarize how a treatment is discussed in the literature, but it should not decide whether a particular person should receive that treatment. The treatment work also provided an important development lesson. Early retrieval was limited because the generic terminology did not adequately represent the therapeutic domain. Reviewing the weak outputs, expanding the vocabulary, and rebuilding the knowledge base produced a more useful graph. This showed that retrieval problems may sometimes be addressed through better domain representation rather than larger models or more complex prompting. Graph relationships connected treatments, disease manifestations, drugs, biomarkers, and reported outcomes. These links supported retrieval but did not independently establish treatment effectiveness. Any interpretation of efficacy remained dependent on the retrieved passage and its source.

Prevention and Counselling Adaptation

The prevention adaptation focused on risk reduction, screening, and counselling rather than specialist procedures. Its database contained 39 papers and 1,640 passages, producing 2,986 entity records and 787 relationships. All passages had stored embeddings, and the final database occupied approximately 16.5 MB. The main topics included glycaemic management, blood pressure control, lipid management, modifiable risk factors, retinal screening, pregnancy related considerations, and general advice after a diagnosis of diabetes. These topics reflect the wider chronic care pathway in which DR prevention takes place. The typical questions included:

- Which modifiable factors are associated with diabetic-retinopathy risk?
- How does glycaemic control relate to progression?
- Why are blood-pressure and lipid management important?
- What considerations may apply during pregnancy?
- Why is regular retinal screening recommended?

The prevention implementation reused the shared processing and retrieval pipeline. Unlike the treatment adaptation, it did not require separately tracking a large dictionary expansion because many prevention concepts were already represented in the general vocabulary. Its main specialization occurred through the selected literature and the counselling-oriented question set. The prevention works also explored differences between retrieval grounded responses and answers from general purpose language models. These comparisons were useful for examining fluency, grounding, and cost, but they were exploration rather than a formal model benchmark. The main lesson was not that one model was universally better. Rather, the retrieval-first framework could expose the passages and concepts supporting its answers, while a general chatbot might provide fluent text without an equivalent evidence trail. This scenario extends DR support beyond ophthalmic procedures. It connects the condition to the broader management of diabetes, including metabolic control, cardiovascular risk factors, timely screening, and pregnancy related monitoring. The framework does not directly manage these factors, but it can organize evidence explaining why they matter.

Comparison of the Two Adaptations

Table 4 presents the main differences between the prevention and treatment adaptations. As shown in Table 4, the treatment database contained more entities and relationships. This does not mean that it performed better than the prevention database. The difference reflects its larger corpus and more detailed terminology for drugs, procedures, biomarkers, and outcomes. The important result is architectural reuse. Both adaptations retained the same core framework and changed only the evidence and knowledge components needed for their intended roles.

Table 4: Adaptation of the shared framework to prevention and treatment scenarios.

Dimension	Prevention Adaptation	Treatment Adaptation
Primary purpose	Risk reduction and counselling education	Treatment and therapeutic evidence access
Main topics	Glycaemia, blood pressure, lipids, pregnancy, screening, and modifiable risks	Anti-VEGF therapy, laser, vitrectomy, drugs, biomarkers, and outcomes
Papers	39	62
Passages	1,640	2,346
Entity records	2,986	6,388
Relationships	787	2,562
Embedded passages	1,640	2,346
Database size	16.5 MB	23.6 MB
Terminology adaptation	Shared base terminology retained	Dictionary expanded from approximately 84 to 190 terms
Example question	How can diabetic-retinopathy risk be reduced?	How is diabetic macular edema treated?
Main lesson	A focused corpus can support counselling-oriented retrieval	Specialized terminology improves treatment-focused retrieval

Worked Example of Prevention Scenario

Figure 4 shows how a prevention-based question moves through the framework: "How can a person with newly diagnosed diabetes reduce the risk of diabetic retinopathy?" The question is allowed because it asks for general educational information rather than a personal diagnosis or treatment decision. The system identifies concepts such as DR, glycaemic control, hypertension, screening, and risk reduction. It then uses semantics, keyword, and KG retrieval to locate relevant evidence. The retrieved material may include passages discussing glycaemic management, blood pressure control, and timely retinal screening. These passages are organized into a concise educational response using the local template or an optional language model. The question passes through safety screening before relevant concepts, graph links, and literature passages are retrieved and presented with source information and an education-only boundary. The resulting answer should explain that DR risk is generally addressed through good diabetes management, attention to modifiable risk factors, and timely retinal screening. It should also be made clear that individual screening schedules and treatment decisions require professional assessment. The example shows the contribution of each part of the system. Safety screening determines whether the question is appropriate, structured concepts guide retrieval, semantic and keyword methods locate supporting passages, and the answer layer presents rather than independently invents the information.

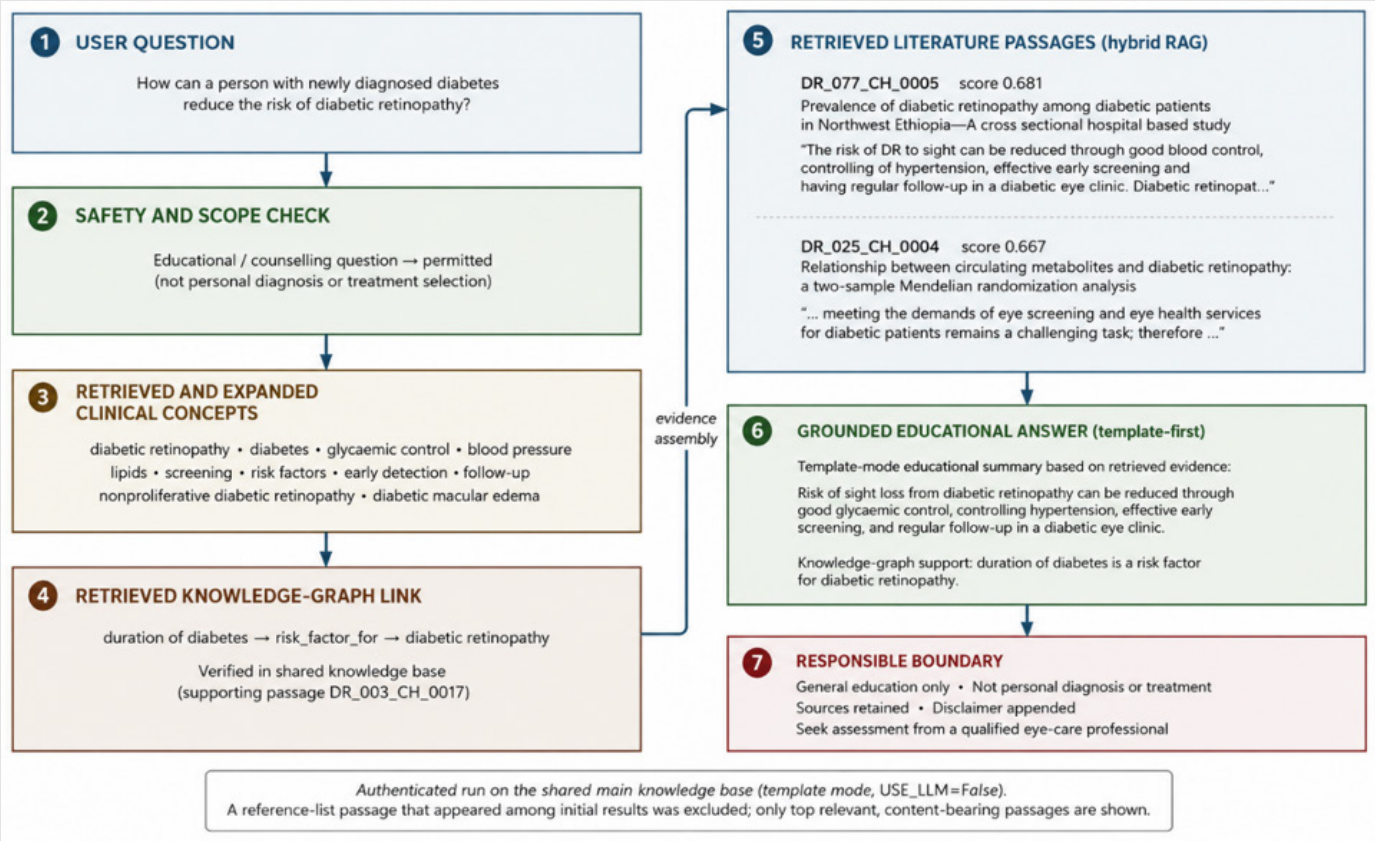

Figure 4: Journey Map from a prevention question to a grounded educational response.

Interpretation and Boundaries

The two adaptations show that a lightweight knowledge support framework can serve different parts of the same chronic care pathway. The treatment version required more detailed therapeutic terminology, while the prevention version relied more heavily on broad risk factors, screening, and counselling concepts. These differences were addressed through knowledge adaptation rather than separate model training programmes. The scenarios also show that structured terminology remains useful even when semantic embeddings are available. Embeddings help retrieve conceptually similar passages, while explicit entities and relationships improve transparency and support focused specialization. The work further shows that important development activities, literature review, terminology expansion, question

design, query inspection, and retrieval assessment can be performed without large-scale computing infrastructure. This creates a practical capacity building opportunity for universities and research groups working under resource limitations. The findings should nevertheless be interpreted as evidence of implementation feasibility rather than clinical effectiveness. The adaptations have not been shown to improve adherence, treatment decisions, visual outcomes, or patient care.

They also differ in corpus size, terminology coverage, and implementation maturity, so direct performance comparisons would require a shared evaluation protocol. Exploratory comparisons with external language models should be interpreted with the same caution. A reliable comparison would require identical questions and evidence conditions, independent reviewers, and predefined criteria for relevance, grounding, clarity, completeness, and safety. Within these boundaries, the two adaptations support the Paper's main claim: one frugal and responsible framework can address distinct diabetic-retinopathy knowledge needs by changing its evidence and terminology layers rather than making large-scale model training the default solution.

Discussion and Conclusions

This PoC shows how DR literature can be transformed into a lightweight and inspectable knowledge support system without placing a large generative model at the centre of the architecture. The framework begins with curated evidence, passage level retrieval, structured clinical concepts, and safety screening. Language generation is introduced only after relevant evidence has been retrieved and can be replaced by a local template when external services are unavailable. The main contribution is therefore not the size of the databases, but the way the system is organized. One shared architecture supported both prevention counselling and treatment evidence access through controlled changes to the corpus, terminology, graph relationships, and question sets. This provides a practical example of how domain specialization can be achieved through better knowledge representation rather than repeated large-model training.

Revisiting The Poc

The shared framework converted selected DR publications into searchable passages, semantic embeddings, clinical entities, and KG relationships stored in a local SQLite database. The same processing and retrieval workflow was then reused for the prevention and treatment adaptations. Three main conclusions can be drawn within the limits of the current evidence. First, a medical knowledge support system can be built using curated literature, compact semantic retrieval, structured terminology, and local storage without training a large generative model. Second, one technical framework can support different educational needs within the same chronic care pathway. As shown in Figure 3 and Table 4, the prevention and treatment adaptations remained part of one architecture rather than becoming separate systems. Third, grounding, transparency, and scope control can be integrated before answer generation. This is more responsible than relying only on a final disclaimer or the apparent confidence of generated text. The database counts in Table 2 show implementation feasibility, but they should not be interpreted as measures of clinical quality. More papers, entities, or relationships do not necessarily lead to better answers. Performance also depends on corpus relevance, text extraction quality, passage boundaries, terminology coverage, retrieval ranking, and the accuracy of the supporting evidence.

Frugality, Responsibility, and Knowledge Adaptation

Frugality in this framework is a system-level property rather than a claim based on one compact model. It results from combining MiniLM embeddings, local SQLite storage, reusable processing scripts, and template-first answer construction. The compact semantic model is used only for retrieval. It does not diagnose disease, select treatments, or generate independent clinical recommendations. Passage embeddings are created during database construction and reused at query time, reducing repeated computation. The databases are also small enough to store, transfer, and inspect ordinary computing systems without a dedicated server. Template-based answering further reduces dependence on commercial language-model services. Evidence can be retrieved and presented locally, while an external model remains an optional wording layer. This separation reduces reliance on continuous connectivity, API funding, and a specific model provider. However, frugality has not yet been fully quantified. The architecture and database sizes show that the system is practically lightweight, but formal measurements of runtime, memory use, response latency, energy consumption, and cost per query are still needed. The framework should therefore be described as architecturally frugal rather than as the result of a complete efficiency benchmark. Responsibility is supported through four main features:

- Bounded literature collection
- Traceable passage retrieval
- Inspectable terminology and graph relationships
- Safety checks applied before ordinary answering.

These mechanisms do not guarantee that every response will be correct, but they make system behaviour easier to examine. When retrieval is weak, the problem can be traced to the corpus, PDF extraction, missing terminology, graph rules, or ranking logic. The safety layer distinguishes educational questions from requests for personal diagnosis, treatment selection, medication changes, or unsupported predictions. This boundary is more meaningful than adding a disclaimer after an unsafe answer has already been produced. The current controls are nevertheless pattern based. They may miss indirect or unusual wording and may also refuse some valid educational questions. Urgent symptoms such as sudden vision loss require a stronger pathway that directs users toward immediate professional care rath-

er than returning a detailed literature summary. Future safety evaluation should therefore measure both missed unsafe requests and false refusals. The treatment adaptation also shows why knowledge level refinement matters. Early retrieval was limited because the shared dictionary did not fully represent drugs, procedures, biomarkers, and treatment outcomes.

Expanding the terminology from approximately 84 to 190 terms improved treatment-focused coverage without changing the embedding model or rebuilding the overall architecture. This indicates that weak retrieval does not always require a larger model. It may result from missing terminology, incomplete relation types, or an evidence collection that is too broad for the intended task. Improving the knowledge layer can therefore be more transparent and easier to maintain than increasing model scale or using increasingly complex prompts. The prevention adaptation required less terminology expansion because many risk factor and screening concepts were already present in the shared vocabulary. Its specialization occurred mainly through a focused corpus and counselling-oriented questions. The two scenarios show that adaptation can occur through corpus selection, controlled terminology, entity categories, relationship rules, and question design. Semantic embedding and structured knowledge therefore serve complementary roles. Embeddings help retrieve conceptually related passages, while explicit entities and relationships improve transparency and support targeted specialization.

DR is also a suitable chronic care use case because it connects long-term diabetes management with risk reduction, screening, disease progression, and treatment. The framework can help users understand these links, but it does not directly manage patients. It does not monitor health data, assess retinal images, prescribe medication, schedule screening, or change treatment plans. Its role is limited to the knowledge and education layer of chronic-disease management.

Global South Relevance and Capacity Building

The framework responds to conditions that are common in resource-constrained health and educational environments, including limited computing capacity, variable connectivity, restricted access to specialists, and limited funding for commercial artificial intelligence platforms. Its use of open tools, local storage, compact retrieval, reusable scripts, and optional cloud generation may make it easier for universities and smaller institutions to build and maintain their own evidence resources. Once a database has been prepared, retrieval and template-based answering can operate locally. Institutions can also review the included publications, add local guidance, and decide whether to send any information to an external service. The modular development process also supported capacity building. The researchers can contribute to literature review, terminology development, database construction, question design, retrieval inspection, and answer comparison without needing to train a large neural model.

This offers a more meaningful introduction to responsible artificial intelligence than using commercial chatbots alone because participants can see how evidence quality, terminology, retrieval, and safety affect the final response. The work should nevertheless be described as GS-based, not as a validated GS deployment. It was not evaluated across countries, clinics, languages, or health systems. Its mainly English language and open-access corpus may also underrepresent regional evidence. Future versions should therefore include local guidelines, regional studies, and multilingual educational resources.

Limitations and Future Priorities

The PoC remains preliminary. Table 5 summarizes its main limitations and corresponding priorities. The highest priority should be a shared evaluation protocol. Independent reviewers should score general, prevention, and treatment questions for relevance, evidence grounding, clarity, completeness, source support, and appropriate scope. Inter-rater agreement should also be reported. A separate safety set should test personal diagnosis, treatment selection, medication changes, urgent symptoms, unsupported predictions, and indirect wording. Both correct refusals and false refusals should be measured. Evidence traceability should also be strengthened. Each response should include stable passage identifiers and clear source information. A locally deployable generative small language model could further reduce dependence on external services, but it should be evaluated only as a wording layer after retrieval, not as an unrestricted medical assistant. Other priorities include multilingual development, locally relevant guidelines, expert-reviewed graph relationships, lightweight user interfaces, and formal version control for the corpus, dictionaries, and database builds. Only after these steps should the framework progress to supervised usability testing with health professionals, educators, or students. Any patient-facing use would require additional ethical review, governance, privacy assessment, and safety validation.

Table 5: Main limitations and future development priorities.

Current limitation	Implication	Priority improvement
Open-access and journal-based selection	Relevant studies, guidelines, and regional evidence may be omitted	Broaden evidence quality criteria and add authoritative guidance
Mainly English-language corpus	Limits accessibility and local relevance	Add Arabic and other locally relevant languages
PDF extraction noise	May weaken passages and retrieval	Improve document validation, cleaning, and table handling
Dictionary-based entity extraction	May miss synonyms, abbreviations, and new terms	Expand expert reviewed terminology and assess clinical named entity recognition
Co-occurrence and rule-based graph links	Relationships do not establish causality or effectiveness	Introduce clearer, expert validated relation types
Limited shared evaluation	Answer quality and retrieval reliability remain uncertain	Create an expert scored benchmark covering general, prevention, and treatment questions
Pattern-based safety controls	Unsafe wording may be missed, while valid questions may be refused	Test diverse safety prompts and strengthen urgent care routing
Optional cloud generation	Introduces privacy, cost, connectivity, and reproducibility concerns	Evaluate a locally deployable generative small language model
Incomplete efficiency measurements	Frugality remains partly qualitative	Measure runtime, memory use, latency, energy, and cost
No clinical deployment	No evidence of improved knowledge, behaviour, or outcomes	Conduct supervised usability studies after technical validation
No formal updating process	The evidence base may become outdated	Add corpus versioning, update logs, and scheduled evidence reviews

Concluding Remarks

This Paper presented a frugal, responsible framework for converting diabetic-retinopathy literature into an accessible, inspectable knowledge-support system. The framework combines curated evidence, passage-level retrieval, a lightweight knowledge graph, compact semantic embeddings, local SQLite storage, safety screening, and template-first answer construction. An external language model may improve wording, but it is not required for retrieval, evidence organization, or source presentation. The prevention and treatment adaptations show that one architecture can support different chronic-care knowledge needs through controlled changes to evidence and terminology. This provides a practical model of specialization in which corpus quality and domain knowledge are strengthened before additional model complexity is introduced. The framework is intended for education and literature navigation, not personal diagnosis, treatment selection, or clinical decision-making. It should therefore be understood as an implementation proof of concept rather than a deployable medical product.

Its main contribution is not an attempt to replace healthcare professionals or compete with the largest language models. Instead, it shows that chronic-disease knowledge support can begin with curated evidence, transparent structures, modest infrastructure, and clear safety boundaries. Although stronger evaluation and real-world testing remain necessary, this approach provides a realistic foundation for locally adaptable and responsible health artificial intelligence in resource-constrained settings.

Acknowledgements

The authors are grateful to Zayed University for a Research and Innovation Grant.

Many thanks are also due to our undergraduate summer research interns – Faisal Alashwal, Sultan Alshehhi and Karthik Ravi – for the data curation (reposited at HYPERLINK "https://protect2.fireeye.com/v1/url?k=46839907-1918a7b8-4687aa08-000babda033e-aeebe2f1a94706c6&q=1&e=9146171f-1b4e-425e-b1b1-fbe7a9112a90&u=https%3A%2F%2Fdrive.google.com%2Fdrive%2Ffolders%2F1mdI8pFX9C3YNRuJaI_48SqpNlqMvXf2F%3Fusp%3Ddrive_link"https://drive.google.com/drive/folders/1mdI8pFX9C3YNRuJaI_48SqpNlqMvXf2F?usp=drive_link) from which the LLM analysis using Claude Opus 5.5 was conducted.

References

1. (2026) IDF Diabetes Atlas. International Diabetes Federation.
2. American Diabetes Association Professional Practice Committee (2024) 12. Retinopathy, Neuropathy, and Foot Care: Standards of Care in Diabetes—2024. Diabetes Care 47(Suppl 1): S231-S243.
3. Z L Teo, Yih Chung Tham, Marco Yu, Miao Li Chee, Tyler Hyungtaek Rim, et al. (2021) Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis. Ophthalmology 128(11): 1580-1591.
4. (2022) Package of eye care interventions. World Health Organization.
5. K Curran, P Piyasena, N Congdon, L Duke, B Malanda, et al. (2023) Inclusion of diabetic retinopathy screening strategies in national-level diabetes care planning in low- and middle-income countries: a scoping review. Health Res Policy Syst 21(1): 2.

6. M M P N Piyasena, Gudlavalleti Venkata S Murthy, Jennifer L Y Yip, Clare Gilbert, Maria Zuurmond, et al. (2019) Systematic review on barriers and enablers for access to diabetic retinopathy screening services in different income settings. PLOS ONE 14(4): e0198979.
7. (2026) Ethics and governance of artificial intelligence for health.
8. (2024) Ethics and Governance of Artificial Intelligence for Health: Large Multi-Modal Models. WHO Guidance World Health Organization.
9. T Ciecierski Holmes, R Singh, M Axt, S Brenner, S Barteit (2022) Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. Npj Digit Med 5(1): 162.
10. J Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, et al. (2019) Do no harm: a roadmap for responsible machine learning for health care. Nat Med 25(9): 1337-1340.
11. P Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, et al. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 9459-9474.
12. W Wang, F Wei, L Dong, H Bao, N Yang, et al. (2026) MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. Advances in Neural Information Processing Systems 5776-5788.
13. N Reimers, I Gurevych (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 3982-3992.
14. E Evangelista, F Ruba, S Bukhari, A Nazir, R Sharma (2026) GraphRAG-Enabled Local Large Language Model for Gestational Diabetes Mellitus: Development of a Proof-of-Concept. JMIR Diabetes 11: e76454.
15. B Wahl, A Cossy Gantner, S Germann, N R Schwalbe (2018) Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?. BMJ Glob Health 3(4): e000798.
16. N Schwalbe, B Wahl (2020) Artificial intelligence and the future of global health. The Lancet 395(10236): 1579-1586.
17. S Kresevic, M Giuffrè, M Ajcevic, A Accardo, L S Crocè, et al. (2024) Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. Npj Digit Med 7(1): 102.
18. C Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, et al. (2024) Almanac — Retrieval-Augmented Language Models for Clinical Medicine. NEJM AI 1(2): 10.
19. G Xiong, Q Jin, Z Lu, A Zhang (2024) Benchmarking Retrieval-Augmented Generation for Medicine. Findings of the Association for Computational Linguistics ACL 2024: 6233-6251.
20. D N Nicholson, C S Greene (2020) Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J 18: 1414-1428.
21. L Wang, Huimin Xie, Wentao Han, Xiao Yang, Lili Shi, et al. (2020) Construction of a knowledge graph for diabetes complications from expert-reviewed clinical evidences. Comput Assist Surg 25(1): 29-35.
22. R Schwartz, J Dodge, N A Smith, O Etzioni (2020) Green AI. Commun ACM 63(12): 54-63.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2026.66.010413

Ravi S Sharma . Biomed J Sci & Tech Res

This work is licensed under Creative *Commons* Attribution 4.0 License

Submission Link: https://biomedres.us/submit-manuscript.php

Assets of Publishing with us

- Global archiving of articles
- Immediate, unrestricted online access
- Rigorous Peer Review Process
- Authors Retain Copyrights
- Unique DOI for all articles

https://biomedres.us/