AI-Powered Retrieval of Medical Literature and Health Data with Vector Databases: Developing a Custom Search Assistant for Medical Research

University of Turku
Department of Computing
Master of Science Thesis
Health Technology
February 2026
Muhammad Junaid Raza

Supervisors:
Tero Koivisto
Jenna Kanerva

UNIVERSITY OF TURKU
Department of Computing

Muhammad Junaid Raza: AI-Powered Retrieval of Medical Literature and Health Data with Vector Databases: Developing a Custom Search Assistant for Medical Research

Master of Science Thesis, 57 p.
Health Technology
February 2026

The rapid growth of online information makes locating precise and relevant biomedical literature increasingly difficult. Researchers often spend significant time filtering useful content from large volumes of unrelated material, making the process inefficient and demanding, particularly in scientific research.

This thesis presents the development of a Retrieval-Augmented Generation (RAG) system for biomedical literature retrieval. The system evaluates three models, GPT, SBERT, and BioBERT, on biosignal-related topics such as ECG, EEG, and PPG. Queries are divided into two categories: expert-formulated and layman-formulated questions. The objective is to assess how accurately each model interprets and answers these different query types. In addition to literature retrieval, the system retrieves and visualizes relevant biosignal datasets, extracted from PhysioNet, to assist researchers in both understanding studies and exploring associated data. Research papers are extracted from PubMed and stored in a database along with embeddings for efficient access.

Both quantitative and qualitative evaluations were conducted. Quantitative performance was initially evaluated using DOI and dataset matching; however, as these metrics were insufficient, embedding-based similarity, TF-IDF similarity, and keyword overlap were added to enhance the evaluation framework. Qualitative performance was assessed using an LLM-as-Judge framework. A custom test set was created containing ground-truth research papers, expert and layman queries, and corresponding dataset URLs from PhysioNet.

Results indicate that GPT achieved the highest answer quality (overall 0.88) and strongest semantic alignment (embedding composite 0.60). SBERT demonstrated nearly identical embedding performance (0.60) and the most stable lexical retrieval performance in TF-IDF (0.50). BioBERT consistently showed lower performance (embedding composite 0.30, TF-IDF 0.17, overall 0.73). Considering the potential evaluation bias toward GPT-family models, specifically, that GPT was used both for final answer generation and for qualitative assessment under the LLM-as-Judge framework, SBERT was selected as the most reliable standalone retrieval model. Using GPT for both generation and evaluation may introduce systematic bias, as models from the same family can exhibit alignment in reasoning patterns and stylistic preferences, potentially inflating performance estimates.

Keywords: RAG, ECG, EEG, PPG, Biomedical Literature, PubMed, PhysioNet, Embeddings, Vector Database, TF-IDF, LLM, LLM-as-Judge

Contents

List of Figures

List of Tables

1 Introduction

Information plays an important role in advancing organizations, and collecting data allows better planning to achieve desired results. Consequently, data collection has become a priority in healthcare to understand current trends and anticipate future developments in patient care and healthcare systems. To address public health demands, new data management and analysis methods must be developed [1]. A divergence exists between new healthcare technology development and its practical application in clinical settings. As digital health technologies and personalized treatment gain popularity, understanding their integration into population-based healthcare is essential. Healthcare organizations aim to combine digital health, artificial intelligence, and precision medicine to improve treatment quality, reduce costs, and address challenges in patient experience and workflow impact [2]. Robust information systems are needed to collect and manage clinical and administrative data effectively.

Hospital Information Systems (HIS), especially in major hospitals, store diverse patient data, including medical records and test results. Although primarily repositories for clinical and administrative data, these systems can support Clinical Decision Support (CDS). Integrating CDS into HIS provides healthcare professionals with timely insights such as risk scores, treatment suggestions, or alerts from ongoing data collection, strengthening clinical decision-making and improving patient care [3]. CDS assist healthcare professionals in making informed clinical decisions,

helping clinicians provide efficient and effective treatments based on patient data. Modern CDS systems incorporate automated risk assessments, medication interaction alerts, and evidence-based recommendations from demographic, clinical, and digital health data. These tools should streamline workflows by integrating and prioritizing critical information. To optimize clinical outcomes, CDS should operate smoothly across healthcare settings, support real-time application, and adopt emerging technologies with minimal disruption [4].

To improve decision-support systems, it is essential to understand the biomedical data that power them. Biomedical data include molecular and omic data, biomedical imaging, clinical records, and electronic health data [5]. Another category is bio-signals, electrical or physiological signals produced by the body that reflect biological activities. These signals, essential for diagnosis, are classified into action potentials, fast voltage changes such as electrocardiogram (ECG), electroencephalogram (EEG), and electromyogram (EMG) [6] and event-related potentials (ERPs), tiny voltages produced in response to events or stimuli, phonocardiogram (PCG) (heart sounds), and carotid pulse (pressure waveform). [6], [7]. Among these biosignals, photoplethysmography (PPG) is a noninvasive optical method that measures blood volume changes in the microvasculature, providing information on cardiovascular, respiratory, and autonomic activity. PPG is widely used to monitor pulse rate and blood oxygen saturation (SpO_2) and can also provide insights into respiratory rate, blood pressure, and other physiological indicators [8].

To summarize, this shift toward data-driven healthcare highlights the need to understand various types and sources of biomedical data. The next section outlines the background of these data and their role in enhancing clinical decision-making.

1.1 Background

Healthcare involves the prevention, diagnosis, and treatment of various medical problems in humans. It includes medical professionals such as doctors, nurses, and allied health workers who operate in clinics, hospitals, and other institutions. These professionals represent different fields, including psychology, physiotherapy, nursing, dentistry, and medicine. Primary care serves as the first point of contact, secondary care provides specialist therapy, tertiary care performs advanced treatments, and quaternary care manages highly complex interventions. Healthcare providers at all levels manage large volumes of data, including test results, medications, diagnoses, medical histories, and other private health information [1]. Traditionally, patient medical records were preserved as handwritten notes or typed reports [1], [9].

However, as the volume and complexity of healthcare data increased, the limitations of paper-based systems became evident. Electronic Health Records (EHRs) gained popularity as a result of the transition to digital systems enabled by advances in computing technology. The Institute of Medicine first introduced clinical information systems (EHRs) in 2003 [1]. The goal was to collect, store, and present electronic health data over time to improve patient care [10]. EHRs now serve as the foundation for data-driven healthcare solutions, including Clinical Decision Support (CDS) systems. However, the continuous growth in both the amount and variety of healthcare data presents new challenges and opportunities for data management and analysis.

1.1.1 Increasing Volume and Complexity of the Data

With the broad adoption of Electronic Health Records (EHRs), the volume and diversity of healthcare data have increased substantially. These data are generally categorized into three types: structured, semi-structured, and unstructured. Structured data, stored in fixed-format databases, include patient demographics,

medications, allergies, and vital signs, while semi-structured data, often presented in flowchart-like formats, consist of elements such as names, values, and timestamps [11]. The emergence of big data has further amplified the scale and heterogeneity of biomedical information, transforming scientific research by accelerating discoveries and enabling data-driven decision-making [12]. However, this growing amount and complexity of healthcare data, driven largely by EHRs, also pose major challenges in data management, integration, and analysis [1], [13].

Unstructured data, the third category, refers to narrative text such as clinical notes, surgical and discharge reports, radiology findings, and pathology records [11]. Although these records are rich in medical information, they lack a consistent format and often contain spelling errors, ambiguous language, and non-standard grammar or punctuation, making them difficult to process and analyze [14]. Clinicians frequently rely on free-text documentation to describe complex conditions, yet this reliance complicates secondary uses of data such as research and analysis [15], [16]. As cited in [17], inconsistencies arise from the interchangeable use of terms, acronyms, and abbreviations. Additionally, a lack of familiarity with medical terminology can hinder understanding and proper use of health information [18].

Zeng et al. [19] demonstrated these challenges by analyzing search queries from patients and healthcare professionals on the Brigham and Women's Hospital website. Their findings revealed notable disparities, including more misspellings in patient queries, lower mapping success to the Unified Medical Language System (UMLS), and differences in semantic category distributions, leading to poorer information retrieval performance for patient queries. Altogether, the predominance of unstructured, free-text data—combined with the ever-growing data volume—intensifies the difficulty of effective information retrieval. Traditional lexical approaches based on the bag-of-words model still dominate most retrieval systems but often suffer from semantic gaps and vocabulary mismatches [20], motivating the need for more ad-

vanced retrieval techniques.

1.1.2 Challenges in Keyword-Based Data Retrieval

Retrieving relevant clinical and medical information through keyword-based search methods presents several significant challenges:

- **Semantic gap**: A disconnect between how medical concepts are expressed in text and how users interpret them, exacerbated by complex medical terminology, acronyms, spelling variations, negations, and temporal references.
- **Vocabulary mismatch**: Search terms used by non-experts or patients often differ from those found in clinical documentation or academic literature, leading to suboptimal search results.
- **Assessing relevance**: Determining the relevance of retrieved information remains difficult due to variations in users' medical knowledge, ambiguous language, and contextual nuances [20].

These challenges can significantly affect retrieval performance. For instance, in the biomedical search engine PubMed, semantically similar queries such as *chlorthalidone vs hydrochlorothiazide* and *chlorthalidone versus hydrochlorothiazide* yield markedly different numbers of relevant articles, with "versus" returning 2.5 times more relevant results [21]. Differences like these illustrate the limitations of traditional term-based search engines that interpret natural language queries as mere collections of terms [22].

The language of biomedical texts is inherently complex, featuring synonymy (different terms with the same meaning), polysemy (terms with multiple meanings), hypernymy (general terms), and hyponymy (specific terms). To address these challenges, biomedical ontologies such as BioMedPlus and controlled vocabularies like MeSH (Medical Subject Headings) can be leveraged to provide synonyms and expand

queries. However, such expansion can exacerbate polysemy, requiring additional techniques to resolve ambiguity. Concept extraction, which derives representative terms from text to generate query sets and analyze retrieved documents, has been applied to rank results according to estimated relevance and mitigate these issues [23].

Overall, keyword-based retrieval in the biomedical domain faces limitations from semantic gaps, vocabulary mismatch, and natural-language complexity. These challenges have motivated the exploration of AI-driven approaches, including vector embeddings, to better capture semantic meaning and improve retrieval effectiveness, which is discussed in the next section.

1.1.3 Role of AI and Vector Embeddings in Biomedical Data Retrieval

Biomedical data, including electronic health records, biosignals, and scientific publications, is growing rapidly, making effective retrieval increasingly challenging. Traditional keyword-based search often fails to capture the semantic meaning of queries and documents, which can limit access to relevant knowledge. To address this, AI-driven methods such as word or document embeddings have been proposed. Embeddings represent text in high-dimensional vector spaces, capturing semantic relationships between terms and enabling more effective retrieval of relevant biomedical information [24]. This high-level approach highlights the opportunities for improving biomedical literature search, while detailed methods and evaluations are discussed in the Literature Review section.

1.2 Problem Statement

The ever-growing volume of biomedical publications makes identifying relevant information increasingly challenging. Literature search, the process of retrieving scientific articles to meet specific information needs, is essential for advancing biomedical research and supporting patient care [25]. Physicians increasingly rely on bibliographic databases to guide patient care, but often face difficulties due to limited time, complex search strategies, and restricted access to some resources [25], [26], [27], [28]. PubMed, a widely used free biomedical database developed by the National Center for Biotechnology Information (NCBI), currently houses over 36 million articles, with more than 1 million new publications added each year [25]. It primarily handles short keyword-based queries and returns lists of raw articles without further analysis, limiting its ability to meet specialized information needs, such as during rapid publication surges like COVID-19 [29], [30], [31], [32], [33].

Traditional keyword-based searches struggle with query ambiguity, fail to capture semantic relationships, and may return irrelevant results [34]. Sorting by publication date can push irrelevant results to the top, and broad or highly specific queries often require manual screening or prior knowledge of exact keywords. Approaches like broad initial queries, rule-based keyword matching, synonym expansion, manual abstract review, and leveraging cited texts offer partial improvements but are limited in scalability and effectiveness [35], [36]. Even with relevance-based sorting and AI-assisted ranking, navigating large result sets remains challenging [25], [37], [38]. Users' heterogeneous knowledge, evolving terminology, and differing methods of reporting further complicate retrieval [39], [40]. Semantic annotations, concept-based filters, and advanced visualization techniques help explore associations between concepts, but efficient retrieval of meaningful results is still difficult.

To address these challenges, AI-driven approaches that leverage semantic embeddings and model-based retrieval offer promising solutions. This thesis investigates

the application of embedding models such as ChatGPT embeddings, SBERT, and BioBERT to enhance the retrieval of biomedical literature and open-source biosignal datasets, providing context-aware, relevance-ranked results. The following research questions and objectives guide the development and evaluation of such a system.

1.3 Research Questions and Objectives

The system evaluates three embedding models—ChatGPT embeddings, SBERT, and BioBERT, to build an AI-powered custom search engine for biomedical literature and biosignal datasets. The database primarily includes research papers on ECG, EEG, and PPG, as well as open-source biosignal datasets, but also contains other biomedical and non-relevant papers to evaluate the system without bias. Each model generates embeddings for papers using the title and abstract, and for datasets using the title and description, storing them in respective columns.

Users can select a model for querying. The system embeds the query using the selected model, matches it against stored embeddings, and returns the most relevant results. For papers, the output includes a summary and DOI link; for datasets, relevant datasets are returned along with a sample plot illustrating the data.

The research questions guiding this thesis focus on AI- and embedding-based approaches for biomedical literature and dataset retrieval, handling terminology variations, and supporting both expert and layperson users:

- **RQ1:** How do different embedding and language models (ChatGPT embeddings + GPT chat model, Sentence-BERT, and BioBERT) compare in terms of retrieval accuracy, relevance, and interpretability for biomedical literature?
- **RQ2:** How well do the models bridge terminology differences between layman and expert queries when retrieving relevant literature and datasets (e.g., “AFib,” “atrial fibrillation,” “fast heart rate”)?

- **RQ3:** How effectively can the system support laypersons in querying biomedical literature compared to domain experts?
  - **RQ3.1:** Does the system produce consistent and accurate answers when the same concept is queried by a layperson vs. a clinical researcher?
- **RQ4:** How effectively can the system retrieve relevant biomedical datasets (e.g., from PhysioNet) based on user queries?

The system focuses on ECG, EEG, and PPG, the most widely studied and relevant biosignals, but also includes other biomedical and non-relevant papers to evaluate performance without bias. Only online, open-source datasets are included. It does not cover other biosignals or health domains and does not provide diagnostic or clinical recommendations, ensuring its purpose is limited to research and educational support.

In summary, this thesis presents an AI-powered search engine evaluating three embedding models for retrieving biomedical literature and open-source biosignal datasets. The system stores model-generated embeddings for papers and datasets, enabling context-aware queries and relevance-ranked results. Research questions focus on model performance, handling terminology variations, supporting both expert and layperson users, and retrieving relevant datasets, while maintaining a scoped and non-clinical focus.

1.4 Thesis Content Summary

This thesis addresses the challenges of retrieving relevant biomedical literature and datasets and investigates how AI and embedding models can improve search effectiveness.

Chapter 1, Introduction, describes the growing volume and complexity of bio medical data, highlighting the limitations of traditional keyword-based searches for

clinicians and researchers. It introduces AI, particularly embedding-based models, as a means to enhance search relevance and concludes with the research problem, objectives, questions, and scope.

Chapter 2, Literature Review, provides an overview of existing biomedical retrieval approaches, contrasting keyword-based search with semantic search techniques. It examines embedding models such as ChatGPT, BioBERT, and Sentence-BERT, discusses vector databases for efficient retrieval, evaluates existing clinical search tools and their limitations, and positions the proposed system as a semantic retrieval framework integrating LLMs and VDBs for enhanced biomedical literature search and knowledge access. This chapter identifies research gaps addressed by the current study.

Chapter 3, Methodology, presents the design and development of the AI-powered search system. It provides an overview of the data sources, including biomedical literature and biosignal datasets, the preprocessing applied to ensure data quality, the generation and storage of embeddings, and the implementation of search strategies using semantic, keyword, and hybrid methods. The chapter also outlines the evaluation framework, including test queries, labeled datasets, and performance metrics.

Chapter 4, System Design, presents the architecture and technical implementation of the proposed biomedical retrieval system. The chapter details how data flows through the system, from ingestion of research articles and biosignal datasets to embedding generation, semantic retrieval, and context-aware answer generation and also discusses the technology stack behind the system's development.

Chapter 5, Results, presents the evaluation of the developed system for answering biomedical questions. It reports the performance of the models across the test set, using multiple evaluation approaches. The chapter also highlights model performance through combined and query type specific results, illustrates cases where

strict DOI matching is insufficient, and discusses both quantitative and qualitative findings to provide a comprehensive overview of each model's strengths and limitations.

Chapter 6, Discussion & Conclusion, interprets the findings in light of the research questions, highlights model performance and query differences, addresses limitations, and situates the results within biomedical research and AI-driven clinical workflows.

1.5 Use of Artificial Intelligence Tools

In the preparation of this thesis, artificial intelligence (AI) tools have been used to support mainly rephrasing and literature search tasks. ChatGPT was employed for rephrasing and grammar checking, as well as for extracting keywords from "hot topics" identified for each biosignal (ECG, EEG, and PPG). These keywords were then used to retrieve relevant papers from PubMed via API. QuillBot and Grammarly AI were used for additional rephrasing suggestions and cross-verifying grammar corrections. All other content in this thesis including interpretations, analyses, and methodologies was developed independently and reflects my own work and ideas. Regarding coding, AI tools were used solely for brainstorming and exploring or comparing potential approaches prior to implementation. All final code, designs, and implementations are the result of my own work.

2 Literature Review

This chapter provides a comprehensive review of literature relevant to the thesis topic, **AI-Powered Retrieval of Medical Literature and Health Data with Vector Databases: Developing a Custom Search Assistant for Medical Research.** It begins by discussing commonly used information retrieval techniques, comparing traditional keyword-based methods with semantic search. The chapter then introduces biomedical embeddings and their applications, gradually narrowing the focus to the specific models employed in this system, including ChatGPT's embedding and chat completion model, BioBERT, and Sentence-BERT. It also examines the storage and retrieval of embeddings via vector databases, highlighting their role in enhancing search performance. Existing clinical search tools are reviewed, with their limitations outlined and connections made to how the proposed system addresses these gaps. Finally, the chapter explores AI applications in biosignal retrieval, emphasizing their increasing importance in medical research and data analysis.

2.1 Traditional vs Semantic Search in Biomedical Literature

This section introduces the two primary information retrieval (IR) techniques: traditional search and semantic search, starting from a general IR perspective and

narrowing down to biomedical literature. It outlines the principles, strengths, and limitations of each approach and presents a comparison highlighting their key differences and relevance to modern biomedical information retrieval systems.

2.1.1 Traditional Search

Information retrieval underpins systems such as search engines, chatbots, and digital libraries, focusing on locating and ranking information that best matches a user's query. Early IR systems, also known as keyword-based or lexical search, relied on exact matches between query and document terms, often using rule-based Boolean logic or domain-specific heuristics. While effective in structured databases, these methods lacked semantic understanding and user intent modeling [41], [42].

Statistical models such as bag-of-words and Term Frequency–Inverse Document Frequency (TF-IDF) improved term relevance estimation [43], but still failed to capture deep semantics. Document ranking assigns relevance scores to documents based on similarity with the query, using methods such as:

- **Boolean models:** Binary classification of relevance by keyword presence.
- **Vector space models:** Represent queries and documents as vectors, ranked by cosine similarity.
- **TF-IDF:** Combines term frequency and inverse document frequency for weighting.
- **BM25:** Probabilistic TF-IDF variant accounting for document length.

Vector space models represent documents and queries as high-dimensional vectors. TF-IDF downweights common terms and emphasizes rarer, more distinctive ones. Most documents contain only a small subset of all possible terms, making vectors sparse and enabling efficient storage and computation in large corpora [44], [45].

Inverted indexes map each term to the documents containing it, enabling fast retrieval without scanning the entire corpus. This is particularly useful for full-text searches. In biomedical IR, inverted indexes are widely used; for example, Elasticsearch leverages them for efficient retrieval, and recent systems have used them to identify patient cohorts based on family disease history [45].

Keyword-based methods, however, struggle with context, negation, and domain-specific semantics. For instance, they may retrieve "cancer" even when a document mentions "no family history of cancer." Such limitations are especially critical in biomedical literature, highlighting the need for semantic search methods [46], [47].

2.1.2 Semantic Search

Semantic search leverages natural language processing (NLP) and machine learning to interpret user intent and document context rather than relying solely on keyword overlap [42], [48]. It improves retrieval accuracy and focuses on meaning while supporting personalization and multimodal data analysis [42]. Applications include contextual literature comparison, knowledge graph construction, and question-answering in medicine. Dense vector representations mitigate vocabulary mismatch by retrieving documents mentioning related terms or synonyms.

Word embeddings represent words as numerical vectors capturing semantic and syntactic relationships [49]. They can be context-independent or contextual, with the latter adapting vectors based on surrounding text [50]. Grounded in the distributional hypothesis, embeddings encode meaning through co-occurrence patterns [51]. Traditional embeddings such as Word2Vec are static and struggle with polysemy and domain-specific terms, motivating contextual embeddings [52]. Contextual embeddings assign each token a vector based on context, dynamically capturing syntactic and semantic properties [53]. For example, "dermatome" may refer to a skin region or a surgical instrument, with different vectors generated depending on usage [50].

Common models include ELMo, BERT, and GPT [54], [55].

Sentence embeddings are fixed-size vectors capturing both sentence content and context. They are critical in biomedical NLP tasks such as information retrieval, semantic search, intent detection, and natural language inference [49]. At the document level, count-based methods such as Latent Semantic Indexing (LSI) compute co-occurrence statistics across term-document matrices and use dimensionality reduction techniques like Singular Value Decomposition (SVD) to project terms and documents into a shared semantic space. Prediction-based methods like Doc2Vec embed documents in the same vector space as words, capturing correlations between words and their containing documents [56]. These approaches enhance document-level comparison and retrieval. Word2Vec remains foundational, positioning semantically related words near each other in a continuous space [57].

Embedding-based methods often complement traditional IR models such as TF-IDF, BM25, and Boolean frameworks, which remain efficient through inverted indexing [58], [59], [60], [61].

2.1.3 Comparison and Summary

Table 2.1 summarizes the key differences between traditional and semantic search in terms of query handling, contextual understanding, and biomedical-specific challenges.

Table 2.1: Comparison of Traditional and Semantic Search

Feature	**Traditional Search**	**Semantic Search**
Handling of synonyms	Poor	Good
Context understanding	Low	High
Biomedical specificity	Limited	Stronger
Scalability	High	Moderate to High

This comparison demonstrates that semantic search overcomes many limitations of traditional keyword-based methods, particularly in interpreting context, handling

synonyms, and managing biomedical-specific terminology. By capturing meaning rather than relying solely on exact word matches, semantic search enables more precise and relevant retrieval in modern biomedical information retrieval systems.

2.2 Overview of Biomedical Embeddings

Biomedical information retrieval (IR) has traditionally relied on keyword-based algorithms like TF-IDF and BM25, which struggle to interpret context or semantic nuance [62]. To overcome these limitations, biomedical text mining (BioTM) has emerged, integrating NLP and AI to extract structured knowledge from unannotated biomedical literature [63], [64]. BioTM combines Information Retrieval (IR), Information Extraction (IE), and Natural Language Processing (NLP) to automate curation, hypothesis generation, and data integration [65], [66].

As BioTM evolved, the need for semantic representation learning became evident. Vector embeddings now serve as powerful features for machine learning models, capturing meaningful relationships between biomedical entities [24], [67]. They have been widely applied to tasks such as named entity recognition and synonym and relation extraction (e.g., chemical–disease, drug–drug, and protein–protein interactions). They are also used for literature retrieval and abbreviation disambiguation.

Wang et al. [24] evaluated embeddings trained on clinical notes, biomedical publications, and general-domain corpora (Wikipedia and news) to assess their biomedical NLP effectiveness. Clinical and biomedical embeddings, trained on Mayo Clinic EHR notes and PubMed Central (PMC) articles, outperformed general embeddings such as GloVe and Google News in both intrinsic similarity tests and downstream NLP tasks.

Transformer-based models further advanced this field through self-attention and encoder–decoder architectures, enabling scalable learning of richer biomedical representations from large-scale data [68].

Deep learning architectures such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) have been widely used for text processing. However, they are limited in modeling long-range dependencies and scalability [69], [70]. Transformers, based on self-attention mechanisms, overcome these limitations by enabling models to focus on contextually relevant parts of text in parallel. In biomedical NLP, transformer models are often initialized from general-domain pretraining and further trained on biomedical corpora for improved domain adaptation [68], [71], [72], [73].

BERT (Bidirectional Encoder Representations from Transformers) introduced bidirectional context modeling using masked language modeling, achieving strong results in named entity recognition (NER), relation extraction, and question answering [74]. However, pretrained on general-domain corpora such as Wikipedia, its understanding of biomedical terminology is limited [75], [76], [77]. BioBERT addresses this by pretraining on large-scale biomedical datasets, including PubMed abstracts and PMC articles, significantly improving biomedical NER, relation extraction, and question answering [75], [78]. Sentence-BERT (SBERT) generates fixed-size sentence embeddings preserving semantic similarity, supporting semantic search, document clustering, and contextual literature retrieval [79], [80].

While BioBERT and SBERT excel in semantic representation, GPT represents a state-of-the-art generative transformer capable of understanding and producing contextually relevant biomedical text, performing tasks such as summarization, information extraction, and interactive question answering with human-like fluency [81]. Beyond generative reasoning, these models provide rich contextual embeddings that can be leveraged for similarity-based retrieval. However, directly querying generative models over large biomedical corpora is computationally infeasible due to the size and complexity of the data. To address this, embeddings extracted from LLMs or domain-specific transformers are stored in vector databases (VDBs),

which are optimized for managing high-dimensional vectors representing semantic and contextual information of biomedical texts [82].

2.3 Vector Databases and Their Role in Embedding Retrieval

Traditional information retrieval relies on sparse, high-dimensional document representations, such as bag-of-words or BM25, which excel at exact term matching but fail to capture semantic similarity [83]. Dense vectors, by compressing information across all dimensions, encode semantic meaning and relationships, enabling searches based on similarity rather than exact matches. This makes dense representations more informative and computationally efficient for tasks such as search, classification, and clustering. A key operation is nearest neighbor search, which becomes costly in high-dimensional spaces due to the "curse of dimensionality." Approximate nearest neighbor (ANN) algorithms address this by retrieving sufficiently close neighbors while balancing accuracy and efficiency. ANN methods reduce the search space via techniques like space-partitioning and proximity graphs such as Hierarchical Navigable Small World (HNSW) algorithm [84] or by dimensionality reduction through quantization and hashing such as Locailty-Sensitive Hashing (LSH) [85], as cited in [86].

To support large-scale ANN search, vector libraries such as Facebook AI Similarity Search (Faiss) [87] implement efficient algorithms e.g HNSW, LSH, often with GPU acceleration for large datasets. While these libraries provide high-performance vector similarity search, they lack full database functionalities required in real-world applications. Vector databases (VDBs) extend these capabilities by combining ANN indexing with database features such as real-time updates, parallel processing, security controls, and integration with analytics and AI tools. VDBs serve as the

backbone for semantic retrieval, caching, and long-term memory in applications such as Retrieval-Augmented Generation (RAG) and LLM-driven workflows [86].

With large language models (LLMs) like ChatGPT, chatbots have become increasingly sophisticated, supporting a wide range of applications. However, LLMs face limitations in semantic retrieval, dynamic knowledge updates, and multimodal data handling. VDBs address these challenges by enabling semantic searches on dense vectors, providing efficient long-term context storage, managing dynamic datasets for incremental learning, and unifying multimodal data for LLM consumption [86].

As transformer-based LLMs continue to advance toward generative and context-aware capabilities, VDBs provide efficient management of high-dimensional embeddings, support semantic search, integrate external knowledge, and mitigate hallucinations [82]. Acting as external memory, they enhance LLM reasoning in RAG workflows by enabling data storage, vectorization, similarity-based retrieval, and context integration [88], [89]. Furthermore, VDBs reduce computational cost and latency by caching embeddings of prior queries [90], [91] and store interaction histories for dynamic context recall [92].

LLMs can also augment database and VDB functionality by providing natural-language interfaces, contextual reasoning, and task adaptability [93]. They can translate prompts into executable queries, optimize performance, enrich incomplete information, and support advanced workflows such as content generation and cross-domain transformations [94], [95], [96], [97], [98], [99], [100], [101]. By integrating prior knowledge and generalizing across database types, LLMs reduce manual effort and improve scalability. While these capabilities highlight the promising role of LLMs in modern data management, a detailed exploration of such integrations is beyond the scope of this thesis.

2.4 Existing Clinical Search Tools and Their Limitations

The rapid growth of biomedical literature offers great potential for evidence-based healthcare, yet existing retrieval systems often underperform due to reliance on exact word matching rather than semantic understanding. Integrating standardized biomedical concepts—such as genes, diseases, and chemicals—has been shown to improve retrieval relevance by capturing relationships beyond lexical overlap [102].

For example, [102] combines sparse retrieval methods (e.g., BM25F) with concept-based extensions like query expansion and re-ranking, and further applies hybrid retrieval where BioMedBERT re-ranks BM25F results to capture deeper semantics. While this enhances accuracy, limitations include high computational cost, incomplete semantic coverage in sparse models, and dependence on frequently updated biomedical concept databases, affecting flexibility and scalability.

Machine learning approaches have also been applied to identify high-quality, clinically relevant literature. Supervised classifiers trained on gold-standard datasets (e.g., ACP Journal Club) improve efficiency but face limitations such as restricted domain coverage, small dataset size, and potential biases. Combining textual features, metadata, MeSH terms, and UMLS concepts can improve performance, and ensemble models often balance recall and precision. However, unsupervised and active learning methods remain largely unexplored, and inconsistent reporting and lack of benchmarks hinder method comparison. Emerging NLP models like BioBERT can further enhance semantic understanding and complement concept-based retrieval [103].

MedCPT, a contrastively pre-trained Transformer model for zero-shot biomedical information retrieval, represents another step toward semantic understanding. Trained on 255 million PubMed query–article click pairs, it integrates a dual-encoder

retriever with a cross-encoder re-ranker to achieve state-of-the-art zero-shot retrieval performance. Despite its strong generalization, MedCPT and similar dense models face challenges of interpretability and may retrieve semantically related but clinically irrelevant results, highlighting the ongoing need for hybrid, interpretable retrieval systems [104].

To address these limitations, the proposed system in this study functions as an intelligent clinical search engine designed for biomedical researchers. It integrates three embedding models: GPT, BioBERT, and SBERT, to enable flexible, model-specific semantic retrieval. Users can input natural-language queries and select their preferred model to obtain summarized answers with corresponding source DOIs. When available, the system also retrieves and visualizes relevant datasets from PhysioNet [105], particularly focusing on ECG, EEG, and PPG signals. By combining multi-model semantic retrieval, evidence summarization, and dataset visualization, the system bridges literature insights with real-world physiological data, addressing both interpretability and accessibility challenges in clinical information retrieval.

3 Methodology

This chapter discusses the methodologies employed in the development of the system, covering the data used, the processes applied to it, and the techniques implemented throughout the workflow.Details of these methodologies, which have been introduced or discussed in previous chapters, are summarized here to provide context for the implementation. The chapter begins with an overview of the data sources, describing the types of data utilized in the system and their respective origins. It then proceeds to outline the steps involved in data ingestion and preprocessing. Following this, it explains the approach used for representing data through embeddings and storing them in a vector database. Subsequently, the chapter presents the search techniques implemented in the system, including semantic, keyword, and hybrid search methods. Finally, it concludes with the evaluation component, providing an overview of the techniques applied to assess system performance and the creation of a labeled test set used as the ground truth for evaluation. The following descriptions provide a high-level overview of the chapter; detailed explanations of each methodology are presented in the respective sections below.

3.1 Data Acquisition and Preprocessing

This section provides an overview of the data sources used in the thesis and also addresses the preprocessing steps applied to prepare the data for later stages of the system. The system relies on two primary types of data: textual data and

biosignal data. The textual data consist of research articles scraped from PubMed, including only the title, abstract, DOI, and open-access articles. The biosignal data include physiological signals, specifically the datasets for ECG, PPG, and EEG, obtained from open-access datasets available on PhysioNet. The following subsections describe each data type in more detail, including the specific datasets used, their relevance to the system, and the preprocessing techniques applied to ensure consistency and quality before further processing.

3.1.1 Textual Data

The textual data used in the system were retrieved from PubMed using the NCBI Entrez Programming Utilities (E-utilities) [106], which provide programmatic access to biomedical literature. To collect relevant papers for each physiological signal, arrays of common or "hot" topics related to ECG, EEG, and PPG were sent as queries to the Entrez API. For each retrieved paper, the system stored the title, abstract, and DOI in a dedicated database; details of how this data is later processed for embedding generation are discussed in Embedding Generation section below.

In addition to these basic fields, a new column called domain_tag was introduced to label each paper according to its corresponding signal. For example, all papers retrieved using ECG-related keywords were assigned the tag ECG, and similarly for EEG and PPG. This labeling facilitates signal-specific retrieval and organization of the textual dataset.

To ensure that the retrieval process was unbiased and to provide a diverse set of examples for evaluation, the dataset also includes irrelevant papers. These irrelevant papers fall into two categories: (1) Biomedical-other, which includes physiological signal papers not related to ECG, EEG, or PPG, as well as other biomedical literature; and (2) Non-relevant, which includes papers from unrelated domains, such as artificial intelligence, computer vision, general science, biology, and agriculture.

The inclusion of these categories allows for a more robust assessment of the system's retrieval performance across relevant and irrelevant texts.

3.1.2 Biosignal Data

The biosignal data used in the system were obtained from PhysioNet, which provides open-access biomedical signal datasets. The PhysioNet API was used to programmatically retrieve available datasets related to ECG, EEG, and PPG signals. Similar to the approach used for textual data, arrays of relevant or "hot" topics for each signal type were sent as queries to collect corresponding datasets.

For each retrieved dataset, the system stored its name, description, URL, and PhysioNet identifier in the database. A domain_tag column was again utilized to assign tags corresponding to each dataset category (e.g., ECG, EEG, or PPG). This tagging facilitates organized storage and signal-specific retrieval during later stages of processing. Details on how this biosignal data is later used are also discussed in Embedding Generation section below.

3.1.3 Data Preprocessing

Data preprocessing was carried out separately for textual data and biosignal data. For the textual data, the preprocessing steps included HTML tag removal, unescaping of HTML entities, removal of non-printable characters, normalization of whitespace, and conversion to lowercase. These steps were applied to the full text of the research papers, including titles and abstracts, as well as to the metadata of the biosignal datasets obtained from PhysioNet, such as titles and descriptions.

Following these cleaning and normalization steps, an additional filtering procedure was applied to ensure data quality. Papers or datasets with missing or excessively short content were excluded from further processing. Specifically, research papers with empty or too-short titles, abstracts, or DOIs were removed, and biosig-

nal datasets with empty or insufficiently descriptive titles, descriptions, or dataset URLs were also filtered out. This ensured that only meaningful and information-rich data were retained for subsequent embedding generation.

3.2 Embedding Generation and Storage

This section describes the techniques used to generate embeddings for both data sources—textual and biosignal data. It outlines the models employed for embedding generation and explains the specific input data used to create embeddings for each modality. Furthermore, it discusses the method and database architecture used to store the generated embeddings for efficient retrieval and similarity search. The approach involved generating embeddings using three different models, and storing them separately so that each model's embeddings could be evaluated and compared in subsequent chapters.

Three models were employed for embedding generation:

- **text-embedding-ada-002 (OpenAI)**: a general-purpose embedding model generating vector representations of 1536 dimensions for text [107].
- **all-mpnet-base-v2 (Sentence-BERT variant)**: a transformer-based model mapping sentences and short paragraphs into a 768-dimensional vector space [108].
- **dmis-lab/biobert-base-cased-v1.1 (BioBERT)**: a domain-specific model pre-trained on biomedical corpora to capture biomedical semantics [109].

For textual data (research papers), embeddings were generated by concatenating the title and abstract of each paper and passing the combined text through all three models. For biosignal data (datasets from PhysioNet), embeddings were generated from the dataset's title combined with its description, using the same three models.

This approach ensures that both textual and biosignal metadata are represented in the same vector-space framework, facilitating unified retrieval.

The embeddings were stored in a vector-enabled database built on PostgreSQL using the PgVector extension. Each model's embeddings are stored in separate tables to allow independent evaluation and comparison. Furthermore, textual-data embeddings and biosignal-data embeddings are stored in distinct tables for modularity and system flexibility. A detailed discussion of the database schema, table names, columns, indexing, and storage strategy is provided in the later chapter System Design.

3.3 Search and Retrieval Pipelines

This section presents the techniques implemented for the search methodology, detailing how users can retrieve answers to their queries. It provides an overview of the search strategies employed within the system, which, as described in previous chapters, allows users to select from three models, GPT, SBERT, and BioBERT, when submitting a query. The retrieval pipeline integrates both semantic search based on embeddings and a hybrid approach combining semantic and keyword-based search techniques depending on the selected model configuration.

The semantic search pipelines, implemented using SBERT and BioBERT models, leverage vector representations of PubMed articles and biosignal datasets. Upon query submission, the system generates an embedding for the query using the selected model. Relevance is then determined by computing the cosine similarity between the query embedding and precomputed embeddings stored in the database. SBERT provides general-purpose embeddings suitable for capturing semantic similarity across diverse biomedical texts, while BioBERT offers domain-specific embeddings trained on biomedical corpora, which enhances retrieval performance for specialized biomedical queries. When a query is submitted, the system generates a

query embedding using the selected model and computes similarity scores against both the article and dataset embeddings. Retrieved items are then ranked based on these similarity scores, and the top-ranked articles and datasets are used to construct the context for subsequent answer generation. In these pipelines, both literature and datasets are retrieved using dense vector similarity, resulting in a fully semantic retrieval framework.

In the GPT-based pipeline, a hybrid search strategy is employed to maximize retrieval relevance. Semantic search is used to identify articles that are conceptually aligned with the query, capturing relevant information even when exact keywords are not present. Keyword-based search is simultaneously applied to specifically retrieve biosignal datasets through pattern matching over dataset names and descriptions rather than vector similarity. This hybrid approach is used in GPT because dataset identifiers often contain acronyms, device names, or protocol-specific labels that are less effectively captured by dense embeddings, whereas SBERT and BioBERT pipelines rely on their respective embeddings for both articles and datasets, ensuring a fully semantic, model-consistent retrieval framework. The results from both searches are integrated into a structured context, which is then provided as input to the GPT model for answer generation. This hybrid configuration reflects the underlying system architecture, where dense embeddings are used for article retrieval, while dataset retrieval relies on lexical matching. The final output delivers a coherent answer that combines relevant article excerpts, dataset descriptions, and, where appropriate, clarifications of biomedical terminology.

3.4 Evaluation

This section presents the methodology employed to evaluate the performance of the proposed search system. The evaluation aims to measure how effectively the system retrieves relevant biomedical articles and biosignal datasets in response to user

queries. To this end, a curated test set comprising 108 queries was created, encompassing ECG, EEG, and PPG signals. Each selected paper includes two questions: one formulated in layman's terms and another in technical clinical language, reflecting diverse user perspectives. For each submitted query, each model retrieves the top $k = 5$ most relevant biomedical articles and the top $k = 5$ biosignal datasets, which are then used to construct the retrieval context.. The evaluation framework combines multiple complementary approaches, including keyword-based assessment, semantic similarity measures, and judgment facilitated by a GPT-based evaluator. Importantly, the evaluation is conducted on the final generated answer produced after the retrieval of articles and datasets, the construction of the combined context, and the subsequent answer generation step, rather than on the retrieved items in isolation. This multi-faceted methodology enables a thorough analysis of the retrieval performance of different models and search strategies, providing a robust foundation for detailed discussion in the subsequent subsections.

3.4.1 Test Queries and Labeled Dataset

The evaluation test set was curated to provide a representative and diverse set of queries across multiple biosignal modalities. Specifically, 40 queries were selected for ECG, 40 for EEG, and 28 for PPG. The selection procedure involved randomly sampling relevant papers from the database, and then, using their titles and abstracts, generating two questions per paper with the assistance of a GPT model: one phrased in layman's terms and another in formal clinical language. Each entry in the labeled dataset contains the paper's title, abstract, DOI, an array of associated questions, and the corresponding dataset URLs. Retaining the title and abstract ensures that even if the retrieved answer does not explicitly reference the original DOI, the evaluation can still compare the generated answer with the source content based on semantic similarity. This structured approach allows for robust

assessment of the system's ability to retrieve relevant articles and datasets across different query formulations.

In constructing these queries, a standardized prompt was used to ensure consistency across all generated entries. After selecting papers from the database for each signal type, their titles, abstracts, and DOIs were passed to the GPT-5.1 model together with a fixed instruction template. The prompt required the model to (i) identify whether the paper was relevant to ECG, EEG, or PPG, (ii) generate two highly specific questions—one formulated in simple layman's language and another in formal clinical terminology and (iii) list any potentially relevant PhysioNet datasets. All fields were returned in a constrained JSON structure to maintain uniformity across the dataset. Although the generation process itself was fully automated, a limited amount of manual curation was performed to verify correct signal-type classification and to exclude outputs that did not adhere to the required format. No content was manually authored; adjustments were restricted to correcting structural inconsistencies or removing entries where the model failed to follow the prompt. The exact prompt used in this process is reproduced below for completeness.

Example Prompt: For every title and abstract that I provide, first determine whether it is relevant to ECG, EEG or PPG. If it is **not** relevant, then only respond: "it is not relevant" and do not generate JSON.

If it is relevant, then produce 2 extremely specific, narrowed-down questions based on that paper's title and abstract:

- One question written in simple layman terms
- One question written in clinical researcher terms

After that, list all potentially related datasets from PhysioNet only. All output should be in small case except the signal type.

The final output must follow exactly this structure:

```
{
"signal_type": "ECG/EEG/PPG",
"paper_title": "",
"paper_abstract": "",
"doi": "",
"dataset": [],
"questions": []
}
```

To illustrate the structure and specificity of the generated entries, one example from each signal modality is provided below.

Example 1 (ECG)

Paper title: *dna damage accumulation impedes cardiac repair after myocardial infarction because of insufficient il-10 expression.* [110]

Abstract: Notably, ku80-deficient mice had reduced anti-inflammatory m2 macrophage infiltration despite exhibiting no significant differences in bone marrow-derived macrophage polarization. in addition, the mrna levels of interleukin-10 (il-10), an anti-inflammatory cytokine essential for m2 macrophage polarization and infiltration, were significantly lower in ku80-deficient hearts than in wt hearts both at baseline and after mi. in situ analysis revealed that cells near the ischemic border zone - likely cardiomyocytes - serve as the major sources of il-10. in vitro studies using hl-1 murine cardiac cells confirmed that chemical hypoxia induces il-10 expression, whereas preexisting dsbs blunt this response. together, these findings suggest that dsb accumulation hinders cardiac repair after mi, potentially because of insufficient il-10 expression in cardiomyocytes, thereby disturbing m2 macrophage recruitment.... [110]

Layman-style question: How does damage to heart cells after a heart attack slow down the healing process in the heart?

Clinical-style question: How does persistent dna double-strand break accumulation impair post-myocardial infarction cardiac remodeling via suppression of cardiomyocyte-derived il-10 expression and subsequent m2 macrophage recruitment?

Relevant datasets: MIT-BIH Arrhythmia Database (MITDB) [111], PTB Diagnostic ECG Database (PTBDB) [112].

Example 2 (EEG)

Paper title: *early assessment and analysis of high-risk factors of neurodevelopmental impairment in neonates with congenital diaphragmatic hernia.* [113]

Abstract: Severe pulmonary hypoplasia (lhr<1.5; or=6.20, 95% ci: 2.15-17.80), pphn, and open surgery are independent predictors of neurodevelopmental impairment in cdh neonates. the combined use of aeeg, rso, and nbna significantly improves the efficiency of early neurodevelopmental impairment identification (auc=0.960), outperforming single indicators. clinicians should prioritize monitoring pulmonary hypoplasia and perinatal complications while adopting multimodal neuromonitoring to optimize early intervention strategies..... [113]

Layman-style question: How can early brain monitoring help predict developmental problems in newborns with a diaphragmatic hernia?

Clinical-style question: What is the predictive value of amplitude-integrated eeg combined with rso and nbna for early identification of neurodevelopmental impairment in neonates with congenital diaphragmatic hernia?

Relevant dataset: CHB-MIT Scalp EEG Database [114].

Example 3 (PPG)

Paper title: *the acute cardiovascular response to dynamic exercise and recovery following cannabis use.* [115]

Abstract: Smoking thc-predominant cannabis elevated post-exercise pulse pressure (pre vs. post; control: 426 vs. 417mmhg, s-thc: 434 vs. 509mmhg, v-thc: 445 vs. 468mmhg, v-cbd: 434 vs. 437mmhg; $p<0.01$). the effect of exercise on arterial stiffness and endothelial function was not modified by cannabis; however, septal isovolumic contraction time (baseline: 7219ms, control: 7620ms, s-thc: 6011ms, v-thc: 6612ms, v-cbd: 6916ms; $p=0.01$) was reduced in s-thc compared to after control exercise ($p=0.048$), indicating altered systolic function. blood pressure during maximal cycling was similar regardless of exposure. systolic function, diastolic function, and ventricular mechanics during exercise were unaffected by cannabis. thc-predominant cannabis increases pulse pressure and alters cardiac, but not vascular, function after exercise. cannabis does not affect blood pressure or cardiac function during exercise..... [115]

Layman-style question: how does cannabis use affect heart pulse pressure and cardiac function after exercise?

Clinical-style question: what are the acute effects of thc- and cbd-predominant cannabis on arterial stiffness, endothelial function, and systolic cardiac parameters during and after maximal dynamic exercise?

Relevant datasets: MIMIC-III Waveform Database [116], MIMIC-IV Waveform Database [117].

3.4.2 Evaluation Metrics

The performance of the retrieval and answer generation system was evaluated using multiple complementary metrics designed to capture both factual correctness and semantic relevance. As described in the previous section, answer generation/generated answer refers to the final response produced after retrieving the top $k = 5$ articles and datasets per query and constructing a combined context, which the model then uses to generate a synthesized answer. Two primary evaluation approaches were employed, providing a balanced assessment of the system's effectiveness.

The first approach, referred to as the **standard test set evaluation**, involved comparing model-generated answers against a curated test set. Metrics included the following:

- **DOI Match:** This binary metric checks whether the generated answer contains the exact DOI of the reference paper. DOIs are a standardized and unambiguous identifier for scientific publications, making this metric a reliable indicator of whether the model correctly retrieved or referenced the intended source. However, DOI matching alone is limited, as a correct DOI does not guarantee that the accompanying explanation is accurate or meaningful.

- **Dataset Match:** This metric evaluates whether the answer includes any relevant dataset URLs from PhysioNet. This metric captures the model's ability to ground its answers in appropriate data sources.

- **Keyword Overlap:** Keyword overlap measures lexical similarity between the generated answer and the reference abstract. This metric provides a lightweight approximation of content alignment and is useful for detecting whether key biomedical terms are present. However, it is sensitive to paraphrasing and does not account for semantic equivalence, which motivates the inclusion of a semantic similarity metric.

- **Embedding Similarity:** Embedding similarity is computed as cosine similarity between embeddings of the answer and the reference abstract using OpenAI's text-embedding-ada-002 model. This metric captures conceptual similarity beyond exact word matching, addressing limitations of keyword overlap in cases of paraphrased or reformulated explanations.

- **TF-IDF Similarity:** TF-IDF similarity is computed as the cosine similarity between TF-IDF vectors of the generated answer and the reference abstract. This metric captures the overlap of important terms weighted by their frequency, providing a measure of lexical similarity that highlights shared keywords and phrases, but does not capture deeper semantic relationships.

- **Composite Score:** To balance the strengths and weaknesses of individual metrics, a composite score was computed as a weighted aggregation of the above metrics to produce a single measure of overall retrieval and answer quality. Computed as:

$$\begin{aligned}\text{composite\_score} = 0.3 \times \text{DOI\_match} + 0.2 \times \text{Dataset\_match} \\ + 0.25 \times \text{Keyword\_overlap} + 0.25 \times \text{Semantic\_similarity}\end{aligned}$$

The second approach, termed **LLM-based evaluation**, is used to address the limitations of strict citation-based metrics. In several cases, model-generated answers did not retrieve the exact DOI present in the test set, yet the responses remained factually accurate and well aligned with the query, indicating that DOI matching alone does not fully capture answer quality. The LLM-based evaluation therefore assesses answer quality relative to the query itself, enabling evaluation of accuracy, completeness, relevance, and citation use even when exact reference identifiers differ, and thereby complementing automatic evaluation metrics.

- **Accuracy:** Evaluates factual correctness of biomedical statements by com-

paring the generated answer against the closest matching reference abstract.

- **Completeness:** Measures the extent to which the answer addresses all relevant aspects of the question, including key mechanisms, entities, and implications described in the reference paper.

- **Relevance:** Assesses whether the answer remains focused on the query without introducing unrelated or misleading information.

- **Citation Use:** Proper inclusion and correct referencing of the DOI(s) and datasets mentioned in the reference paper, penalizing missing or incorrect citations.

- **Overall:** Holistic judgment integrating accuracy, completeness, relevance, and citation use, while also considering clarity, conciseness, and structured explanation.

In practice, the evaluation pipeline first identifies the reference paper with the highest semantic similarity to the query using embeddings, and then GPT is prompted to score the model answers on a 0–1 scale across the metrics above. The model is instructed to compare answers to the reference abstract, DOI, and datasets, penalize hallucinations, reward well-structured and concise responses, and provide brief comments explaining each score. The output is formatted as a JSON array for standardized analysis across models.

This combination of quantitative and qualitative evaluation provides a robust framework for assessing both retrieval accuracy and answer quality, facilitating meaningful comparisons across different retrieval strategies such as semantic-only, keyword-only, hybrid and answer generation models.

4 System Design

This chapter presents the design of the proposed system, detailing both its architecture and the technology stack used for development. It provides a comprehensive description of the main components, including the ingestion pipeline, preprocessing engine, embedding and vector store layer, retrieval API, and the frontend interface. Each component is discussed individually in terms of its role, functionality, and interaction with other modules. Finally, the chapter illustrates how these components are integrated into a cohesive system, highlighting the data flow and overall system architecture with a schematic representation.

4.1 Architecture

This section provides an overview of the system architecture, outlining the main components and their interactions. The architecture is designed to efficiently handle the end-to-end workflow, from data ingestion and preprocessing to embedding, storage, retrieval, and presentation of results through the frontend interface. Each component is described individually in the following subsections, highlighting its role, inputs and outputs, and the technologies employed. Figure 4.1 illustrates the overall system architecture and the relationships between its components.

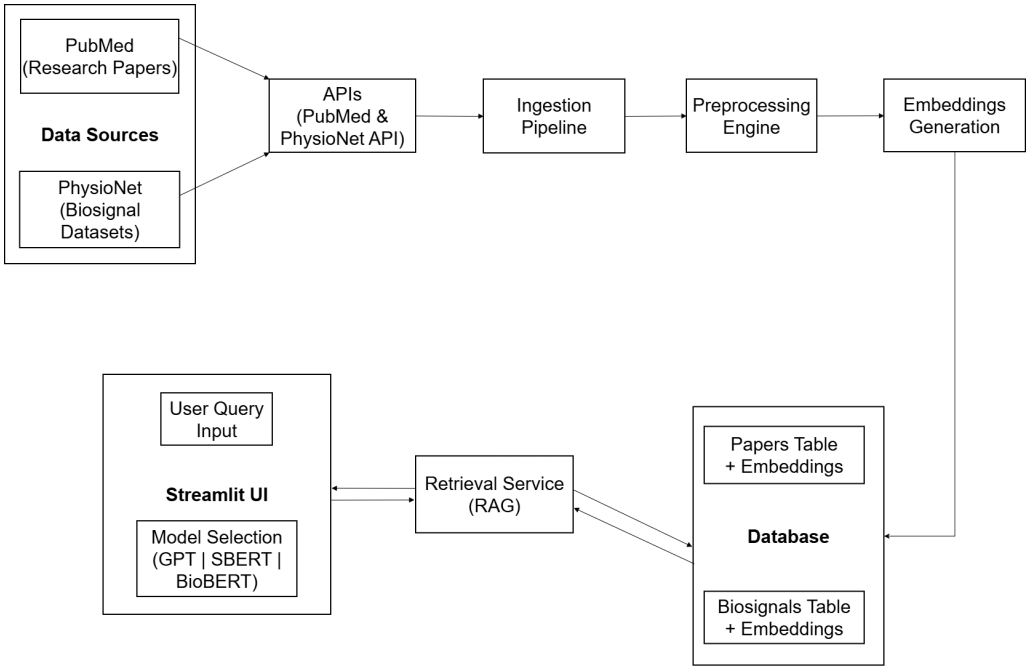

Figure 4.1: System architecture of the proposed framework

4.1.1 Ingestion Pipeline

The ingestion pipeline is responsible for collecting data from external sources and preparing it for downstream preprocessing and embedding. The system ingests two primary data types: (i) biomedical research articles retrieved from PubMed, and (ii) physiological signal datasets obtained from PhysioNet. Since ingestion is treated as an administrative functionality, it runs separately from the user-facing interface and is executed only by authorized system administrators.

The ingestion process begins by querying external data sources using their respective APIs, as previously discussed in subsection Textual Data of the Methodology chapter. Both APIs support keyword-based search, allowing the system administrator to specify terms of interest that determine which papers or datasets are retrieved. For PubMed, the pipeline collects metadata such as the article title, abstract and Digital Object Identifier (DOI). For PhysioNet, the pipeline retrieves metadata including the dataset title, description and URL.

Once the content is retrieved, the ingestion layer forwards the data directly

to the preprocessing engine. At this stage, the data remains in its raw form; no embeddings or storage operations occur before preprocessing. The preprocessing engine is responsible for cleaning and normalizing the retrieved content so that only validated, structured, and high-quality information is stored in the PgVector-enabled PostgreSQL database. This staged design ensures that the database only contains processed and consistent data, enabling efficient downstream embedding and retrieval.

4.1.2 Preprocessing Engine

The preprocessing engine applies all cleaning and normalization steps to the raw data retrieved during ingestion as previously discussed in subsection Data Preprocessing of the Methodology chapter. These transformations ensure that both textual and physiological data follow a consistent schema suitable for downstream embedding. After validation, the ingested data is stored in a PostgreSQL database extended with the PgVector extension. Two separate tables are maintained: one dedicated to research articles and one to biosignal datasets. This separation ensures clean organization of heterogeneous data types and enables seamless retrieval and indexing during the embedding and similarity search stages. Each entry is stored with its associated metadata and a unique identifier, ensuring traceability throughout the system. Upon successful ingestion, the stored data becomes available to the preprocessing engine, which handles cleaning, normalization, and preparation of text and biosignal data for embedding. The modular design of the ingestion pipeline ensures maintainability, extensibility, and consistent data flow into the subsequent stages of the architecture.

4.1.3 Embedding and Vector Store Layer

As explained earlier in section Embedding Generation and Storage of the Methodology chapter, three different models are employed to generate embeddings for the dataset. These models are applied to both research articles and biosignal datasets, producing fixed-dimensional vector representations that capture semantic and structural information. For textual data, the models encode each article abstract and title combined into dense vectors representing its contextual meaning. For biosignals, it combines the title and description of and transforms it into vector form.

The generated embeddings are stored in the PostgreSQL database extended with the PgVector extension. Separate tables are maintained for research articles and biosignals, and embeddings from each model are stored in dedicated columns to facilitate multi-model retrieval. During query execution, user inputs are encoded using the same embedding models, generating query vectors that are compared against the stored vectors. The system then returns the top-k most relevant articles or biosignal datasets according to semantic similarity. This approach enables user to perform meaningful searches based on content rather than simple keyword matches, providing a robust foundation for downstream retrieval and analysis.

4.1.4 Retrieval Service

The retrieval service serves as the core interface between the user and the system, enabling user to submit queries and receive relevant answers from both research articles and biosignal datasets. The user first selects one of the available embedding models. The system transforms the user query into a vector using the chosen model and performs a similarity search against the set of precomputed embeddings in the PgVector database corresponding to that specific model. Based on similarity scores, the top-k most relevant papers or datasets are retrieved.

The retrieved items are then used to construct a contextual prompt for the

ChatGPT completion model. Specifically, the system utilizes the GPT-4-mini model as mentioned earlier in Methodology chapter, to generate answers while following explicit instructions to avoid hallucinations, include DOIs for research articles, and provide dataset URLs for biosignals. The context includes the content from the top-k retrieved articles or datasets, ensuring that the generated answer is grounded in the source material.

For biosignal datasets, the system first encodes the user query into an embedding and performs a top-k similarity search against the precomputed embeddings of all biosignal datasets. The retrieved datasets are then used to build a context, similar to the article retrieval process, which guides the generation of answers. The system subsequently retrieves the dataset URLs and uses the `WFDB` Python library [118] to directly visualize the signals without downloading them to local storage. This approach maintains data integrity, reduces storage overhead, and allows real-time visualization of physiological signals.

The frontend serves as the user interface, enabling them to submit queries, select models, and visualize results; the implementation details are discussed in Section User Interface.

4.1.5 User Interface

The system provides a user-friendly interface for the users through a Streamlit [119] based frontend, allowing interaction with the backend retrieval and embedding pipeline. The interface consists of a text input box for the user to submit queries and a dropdown menu to select the embedding model from which they want the answer to be generated. A submit button triggers the query submission process.

Upon submission, a loading placeholder is displayed while the system processes the query. Once the answer is generated by the backend GPT-4-mini model, the interface presents the response along with clickable Digital Object Identifiers (DOIs)

for research articles and clickable URLs for biosignal datasets. This ensures that users can directly access the original sources referenced in the answer.

If the query references a biosignal dataset, the interface then displays an additional loading placeholder while the system retrieves and processes the dataset. The dataset is visualized directly within the frontend using the WFDB Python library [118], without requiring local downloads, allowing real-time inspection of physiological signals.

The interface design, combined with its integration with the retrieval and embedding backend, provides clinical researchers with an intuitive and efficient platform to explore biomedical literature and biosignal data. A screenshot of the frontend is shown in Figures 4.2 and 4.3.

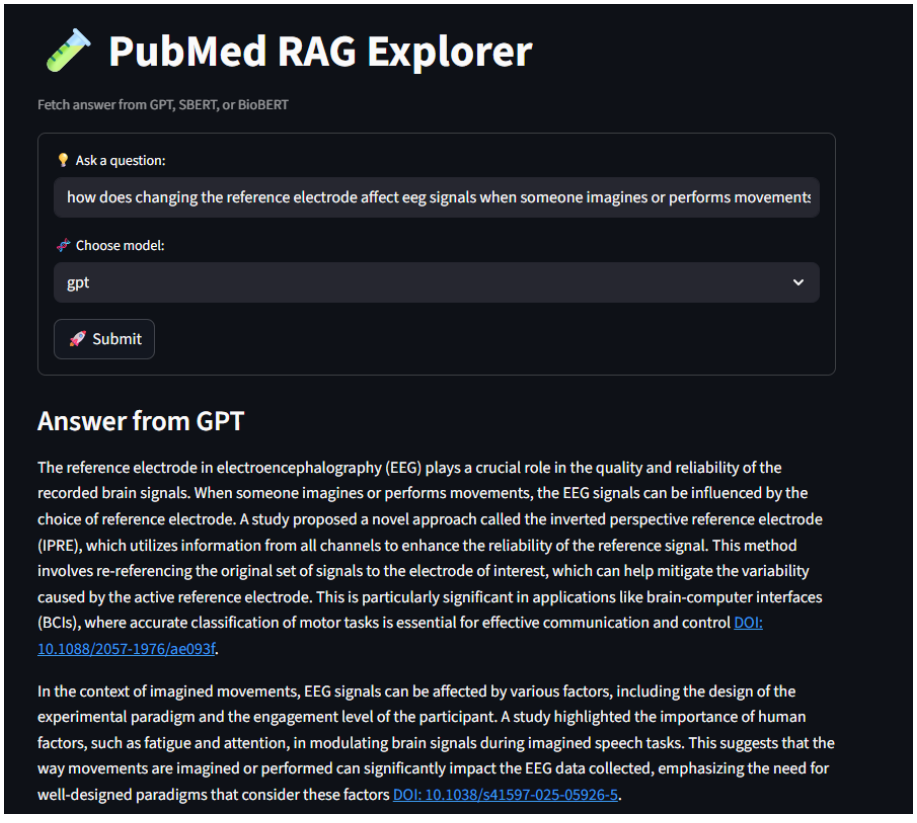

Figure 4.2: Streamlit frontend with query input, model selection, and answer display including clickable DOIs and dataset URLs.

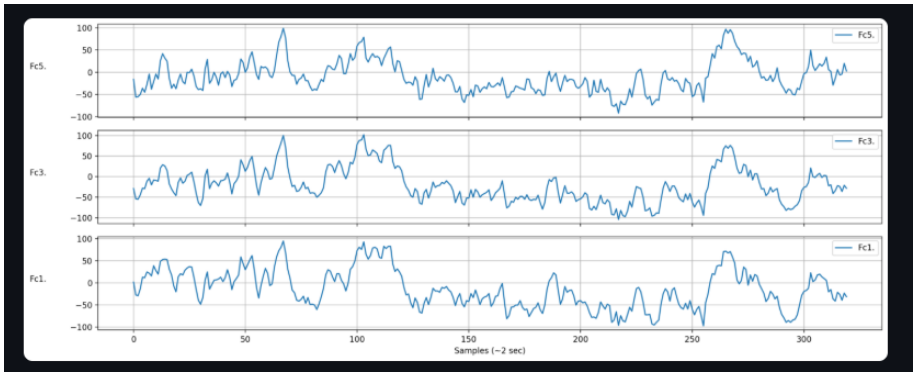

Figure 4.3: Streamlit frontend showing the retrieved biosignal visualization.

5 Results

This chapter presents the evaluation of the developed system for answering biomedical questions. The system is assessed using three models i.e. GPT, SBERT, and BioBERT, across both layman and expert questions. Multiple evaluation approaches were employed, including similarity-based evaluation, which measures alignment using embedding similarity and TF-IDF similarity, as well as assessments conducted through the GPT-as-Judge pipeline. Each approach was applied to both layman and expert questions, separately and combined, to analyze the performance differences and highlight the comparative strengths and limitations of each model. The results are summarized in tables to provide a clear and structured overview of the performance metrics.

5.1 Evaluation Techniques and Results

To evaluate the system, model-generated answers are compared with the test set. The assessment uses semantic embeddings of the answers and reference abstracts, allowing similarity to be measured beyond exact word matches and enabling automated evaluation across question types. The first evaluation step computes DOI match and Dataset match for all questions, including both layman and expert queries.

Table 5.1 presents the combined results using these two metrics. The results show that both GPT and SBERT effectively capture DOI information, with scores

Table 5.1: Layman & expert questions evaluation using Embeddings

Metric	GPT	SBERT	BioBERT
DOI match	0.89	0.88	0.06
Dataset match	0.00	0.06	0.00

of 0.89 and 0.88 respectively, while BioBERT remains low at 0.06, indicating difficulty in correctly referencing sources. Dataset match is minimal for all models, suggesting limited recognition of dataset URLs. This is partly because multiple relevant datasets may exist for a given query, and the system sometimes retrieves a different but still appropriate dataset than the one listed in the test set. Additionally, in some cases no dataset is retrieved, or the retrieved dataset URL does not exactly match the expected test-set entry, even though the content may be semantically correct. As a result, exact-match metrics underestimate the true usefulness of the retrieved datasets. The following example illustrates this situation:

Query: *Can traditional Chinese medicine help people with diabetes and heart disease without affecting heart safety as seen on ECG?*

Model: SBERT

Expected Dataset Title: *MIT-BIH Arrhythmia Database* [111]

Retrieved Dataset Title: *BIDMC Congestive Heart Failure Database* [120]

This reflects a limitation of the dataset itself rather than the retrieval models, and while dataset retrieval could be improved with better annotations, addressing this is beyond the scope of this thesis.

However, it is important to note that DOI and dataset match are strict exact-match metrics. A system may produce a correct and relevant answer, but if the retrieved DOI or dataset does not exactly match the expected one, the result is still marked as incorrect. This limitation shows the need for additional evaluation criteria beyond exact matching.

5.1.1 Similarity based Evaluation

To provide more flexible measures of alignment with the abstract's terminology and content, three similarity metrics were used: keyword overlap, embedding similarity, and TF-IDF similarity. Keyword overlap measures the proportion of words from the reference abstract that are present in the generated answer, providing a simple and interpretable assessment of lexical alignment and ensuring that key terms are explicitly included in the final answer. Embedding similarity was calculated by generating vector embeddings for each model-generated answer and the corresponding reference abstract using OpenAI's `text-embedding-ada-002` model, and then computing the cosine similarity between these vectors. TF-IDF similarity was computed by transforming the answer and abstract into TF-IDF vectors and calculating the cosine similarity between them. These metrics allow evaluation of both semantic alignment and lexical overlap beyond strict exact matches.

Table 5.2: Layman & expert questions evaluation using Embeddings (N = 216)

Metric	GPT	SBERT	BioBERT
DOI match	0.89	0.88	0.06
Dataset match	0.00	0.06	0.00
Keyword overlap	0.38	0.38	0.24
Embedding similarity	0.93	0.93	0.89
TF-IDF similarity	0.51	0.50	0.37
Composite score	0.60	0.60	0.30

Table 5.2 presents the evaluation results for layman and expert questions across the three models. GPT and SBERT show moderate keyword overlap (0.38), indicating reasonable alignment with the abstracts' terminology, while BioBERT is lower (0.24), suggesting weaker lexical correspondence. Embedding similarity is very high for GPT and SBERT (0.93), demonstrating that their answers closely capture the abstracts' semantic content, with BioBERT slightly lower (0.89) but still strong. TF-IDF similarity is lower than embedding similarity for all models (GPT 0.51, SBERT 0.50, BioBERT 0.37), reflecting differences in exact phrasing and highlighting lexical

overlap rather than deeper semantic correspondence. The composite score summarizes these measures, showing that GPT and SBERT perform comparably (0.60), whereas BioBERT is substantially lower (0.30). DOI and dataset match values referenced here come from Table 5.1 and were not newly computed. Finally, results were also analyzed by query type, as each paper includes one layman-formulated question and one expert-formulated question, ensuring evaluation across both user perspectives.

Evaluation of Layman Queries

Table 5.3: Layman questions evaluation using Embeddings (N = 108)

Metric	GPT	SBERT	BioBERT
DOI match	0.89	0.87	0.06
Dataset match	0.00	0.06	0.00
Keyword overlap	0.36	0.37	0.25
Embedding similarity	0.93	0.92	0.89
TF-IDF similarity	0.48	0.47	0.38
Composite score	0.59	0.59	0.30

Table 5.3 shows that GPT and SBERT accurately capture DOI information (0.89 and 0.87), while BioBERT performs poorly (0.06), indicating difficulty in correct source referencing. Keyword overlap is moderate for GPT and SBERT (0.36, 0.37), reflecting reasonable alignment of terminology with the abstracts, whereas BioBERT is lower (0.25). Embedding similarity is highest for GPT (0.93), followed by SBERT (0.92) and BioBERT (0.89), indicating strong semantic alignment with the abstract content. TF-IDF similarity is lower for all models (GPT 0.48, SBERT 0.47, BioBERT 0.38), capturing lexical overlap rather than deeper semantic content. The composite score confirms this pattern: GPT and SBERT perform similarly (0.59), while BioBERT is substantially lower (0.30). Overall, GPT and SBERT are more effective for layman queries in terms of content accuracy and semantic relevance.

Evaluation of Expert Queries

Table 5.4: Expert questions evaluation using Embeddings (N = 108)

Metric	GPT	SBERT	BioBERT
DOI match	0.90	0.90	0.06
Dataset match	0.00	0.06	0.00
Keyword overlap	0.40	0.39	0.24
Embedding similarity	0.93	0.93	0.89
TF-IDF similarity	0.53	0.52	0.36
Composite score	0.60	0.61	0.30

Table 5.4 shows a pattern similar to the layman queries. GPT and SBERT both achieve strong DOI match scores (0.90), while BioBERT remains very low (0.06), indicating limited ability to reference the correct source. Keyword overlap is slightly higher for GPT (0.40) and SBERT (0.39), indicating better capture of expert-level terminology compared with BioBERT (0.24). Embedding similarity remains high for GPT and SBERT (0.93), with BioBERT slightly lower (0.89), reflecting strong semantic alignment with the abstract content. TF-IDF similarity is lower for all models (GPT 0.53, SBERT 0.52, BioBERT 0.36), measuring lexical overlap rather than deeper semantic meaning. Composite scores further confirm this trend: GPT (0.60) and SBERT (0.61) perform consistently well, while BioBERT (0.30) lags behind. Overall, GPT and SBERT provide more accurate, relevant, and semantically aligned answers for expert-formulated queries.

Overall, the evaluation shows that the performance trends for layman and expert queries are highly consistent across all models. GPT and SBERT maintain strong semantic alignment, keyword coverage, and DOI accuracy for both question types, while BioBERT consistently under-performs. This indicates that the choice between layman- or expert-formulated queries has minimal impact on retrieval and answer quality, and either can be effectively used in the search system.

5.2 GPT-as-Judge Evaluation

This section presents the GPT-as-Judge Evaluation, which is employed to complement the quantitative evaluation methods discussed earlier. The primary objective of this approach is to examine cases where the system-generated responses do not exactly match the DOI references in the test set but are nevertheless accurate, relevant, and aligned with the intent of the question. While TF-IDF– and embedding-based evaluations partially address such scenarios through semantic similarity measures, they are limited in capturing qualitative aspects of answer correctness. The GPT-as-Judge framework enables a deeper, context-aware assessment of model outputs by evaluating factual accuracy, relevance, citation use, and overall response quality, thereby providing a more comprehensive evaluation of the system's performance. To better illustrate this evaluation approach, consider the following example query along with its expected answer.

Example question from test set: How does damage to heart cells after a heart attack slow down the healing process in the heart?

Expected Paper Title: *Efficacy and safety of Tongmai Jiangtang Capsule in the treatment of type 2 diabetes mellitus complicated with coronary heart disease with syndrome of damp–heat obstructing collaterals.*[121]

Retrieved Paper Titles:

- **GPT**: *Efficacy and safety of Traditional Chinese Medicine in alleviating symptoms associated with myocardial bridge: a systematic review and meta-analysis.* [122]
- **SBERT**: *Effectiveness and safety analysis of Qifu Yixin Prescription for the treatment of heart failure with preserved ejection fraction: study protocol for a randomized, double-blind, placebo-controlled clinical trial.* [123]

- **BioBERT**: *Changes and monitoring technology of human heart rate and blood oxygen saturation under high-altitude hypoxia.* [124]

Table 5.5: Example query results

Model	Expected DOI	Retrieved DOI
GPT	10.1016/j.phymed.2025.157234	10.3389/fphar.2025.1619617
SBERT	10.1016/j.phymed.2025.157234	10.1186/s12906-025-05106-3
BioBERT	10.1016/j.phymed.2025.157234	10.3389/fphys.2025.1642777

Table 5.5 is based on the example question from the test set described above and illustrates a case where the retrieved DOI does not match the expected DOI, which is treated as the ground truth in the test set. Although the responses generated by all models contain multiple DOI references, only the primary DOI is considered to ensure clarity in the evaluation. This example highlights a key limitation of relying solely on DOI matching to assess system accuracy, as correct and meaningful responses may still fail strict reference-based metrics. To address this limitation, the responses are further evaluated using the GPT-as-Judge framework, in which a large language model assesses the quality and relevance of the generated answers, as described in Section 3.4. The final outputs from all three models—generated, as mentioned earlier in Section 3.4, after retrieving the top $k = 5$ articles and datasets for each query, are submitted to the LLM-as-Judge pipeline to determine the evaluation verdict.

Table 5.6: Evaluation Using GPT-as-Judge

Metric	GPT	SBERT	BioBERT
Accuracy	0.90	0.85	0.75
Completeness	0.85	0.80	0.70
Relevance	0.90	0.85	0.75
Citation	0.90	0.85	0.70
Readability	0.85	0.80	0.75
Overall	0.88	0.83	0.73

Table 5.6 presents the qualitative evaluation results using the GPT-as-Judge framework. GPT outperforms SBERT and BioBERT across all criteria, with high

accuracy and relevance scores (0.90) indicating stronger alignment with reference abstracts and factual correctness. Its higher completeness and citation scores reflect thorough coverage and effective use of references. SBERT scores competitively but lower, showing strength in semantic similarity yet limitations in fully articulated, well-cited answers. BioBERT performs lowest overall, especially in completeness and citation use. Readability scores further highlight generative models' advantage in producing clear, coherent explanations for non-expert users. These results reinforce the limitations observed in Table 5.5, demonstrating that strict DOI-based evaluation alone is insufficient, and that holistic, reference-grounded assessment via GPT-as-Judge provides a more reliable measure of system performance.

In addition to these metrics, the evaluation framework includes a parameter termed "comments", which contains the judge's qualitative feedback on each generated answer. The following section presents the judge's comments for the responses produced by each model.

- **GPT**: *"The GPT model provides a comprehensive overview of TCM's potential benefits for diabetes and heart disease, referencing the relevant study effectively. It discusses ECG monitoring, which is crucial for assessing heart safety. The answer is well-structured and clear, making it accessible to a non-expert audience."*

- **SBERT**: *"The SBERT model effectively summarizes the findings of the TJC study and mentions the importance of ECG monitoring. However, it lacks some depth in discussing the implications of the findings on heart safety. The answer is clear and relevant, but slightly less comprehensive than the GPT model."*

- **BioBERT**: *"The BioBERT model provides a more general discussion about TCM and its potential benefits but does not directly address the specific study on*

TJC or its outcomes. It mentions the importance of ECG monitoring but lacks specific references to the provided studies. The answer is somewhat relevant but less focused on the question than the other models."

The comments parameter provides qualitative feedback from a judge on each model's answer, reflecting content accuracy, clarity, and relevance. GPT's answer was comprehensive and well-structured, SBERT was clear but slightly less detailed, and BioBERT was less focused and relevant. This demonstrates that even when the DOI does not match, qualitative evaluation is necessary to assess answer accuracy based on multiple factors.

5.2.1 Limitations of GPT-as-Judge

While using an LLM as a judge demonstrates that strict DOI matching is not the only valid evaluation criterion, and that the accuracy and relevance of retrieved results should also be considered even when the test set DOI is not matched, this approach has several limitations. One key concern is model-family bias, as the same OpenAI model is used both to generate the candidate answers and to evaluate them. Another limitation is order sensitivity: LLMs can produce different judgments depending on the sequence in which answers are presented. To reduce this effect, the evaluation pipeline randomly shuffles the answers before each run, thereby decreasing the chance of positional bias. Despite these measures, the random nature of LLM outputs means that judgments may still vary slightly across runs, highlighting the need to interpret the results as indicative rather than absolutely deterministic.

5.3 Comparison of Results and Summary

Across all evaluation techniques embedding-based, TF-IDF, and GPT-as-Judge, GPT consistently achieves the highest performance, demonstrating strong alignment

with reference abstracts, higher semantic and lexical correspondence, and superior readability. SBERT performs comparably in embedding and TF-IDF evaluations, particularly when considering models outside the same family, and shows competitive semantic understanding. BioBERT consistently underperforms, primarily due to limited domain-specific training data, resulting in lower completeness, citation use, and semantic alignment.

As mentioned in the previous subsection, it is important to note that the GPT-as-Judge evaluation may favor models from the same family as the judge itself. When accounting for this potential bias, SBERT emerges as the most reliable model in terms of cross-model evaluation. BioBERT's performance could be improved by fine-tuning with domain-specific corpora or incorporating a trained encoder on top of the base model, an approach that can be explored in future work.

6 Discussion & Conclusion

This chapter provides a discussion of the findings presented in the previous chapter and situates them within the context of the research questions. It aims to interpret the results and discuss the implications of the study. The discussion also considers the performance of the system across different query types, identifies limitations of the methodology, and suggests directions for future work. By reflecting on the overall outcomes, this chapter connects the experimental results to the broader objectives of the thesis and provides recommendations for further research and practical applications.

6.1 Interpretation of Results

This section begins by discussing the types of biosignals included in the study, the criteria for selecting hot topics, and the rationale behind these choices. It then provides an overview of the system's general findings, highlighting the performance of each model (GPT, SBERT, and BioBERT) and suggesting which model may be most suitable for this domain or problem. The section also examines the results of layman versus expert queries, offering insights into how well each model interprets different types of questions. All of these observations are based on the detailed findings presented in Chapter 5. Finally, the discussion evaluates how effectively the system addresses the research questions and summarizes the overall insights drawn from the study.

6.1.1 Scope of Biosignals and Topic Selection

The system is built on literature related to three biosignals: ECG, EEG, and PPG. These were chosen because they are well-established biomedical signal domains with abundant publicly available datasets, standardized terminology, and comparable methodological frameworks. This consistency enables controlled evaluation of source grounding, keyword alignment, and semantic understanding in biomedical question-answering systems. At the same time, the three signals originate from different physiological systems, ensuring domain diversity. Within each category, widely researched and high-interest topics were selected to maintain practical relevance and realistic information needs. To avoid evaluation bias, additional papers from other biomedical and non-biomedical areas were also included.

6.1.2 Overview of the System and General Findings

The system is designed around several core objectives that directly correspond to the research questions of this thesis. First, it evaluates how different embedding and language models (GPT-based embeddings with a GPT chat model, SBERT, and BioBERT) compare in terms of retrieval accuracy, relevance, and interpretability, addressing RQ1. This evaluation focuses on how effectively each model retrieves literature and generates responses grounded in the correct sources.

Second, the system examines how well these models handle terminology variation and synonymy, which relates to RQ2. To assess this, the curated test set includes two forms of each query: one phrased in layman language and the other in expert terminology. This design also supports RQ3, which investigates how effectively the system serves both laypersons and domain experts. By comparing model performance across these two query types, the study evaluates answer consistency, terminology understanding, and the system's ability to bridge the language gap between general users and specialists. In addition, the system incorporates visual-

ization of biosignals to improve interpretability and practical usability of retrieved information, supporting the broader objective of enhancing user understanding.

Regarding dataset retrieval, linked to RQ4, results indicate that dataset matching performance is not as strong as literature retrieval. However, this metric specifically measures whether the retrieved dataset exactly matches the gold-standard dataset in the test set. The system is still capable of retrieving relevant datasets, but exact matching is limited because dataset retrieval currently relies primarily on keyword-based methods. More advanced semantic or hybrid retrieval techniques could improve this however, this limitation comes from the dataset itself, not the model. Improving dataset annotations could make retrieval better, but that is beyond the scope of this thesis.

As discussed in Chapter 5, GPT-based models showed the strongest overall performance compared to SBERT and BioBERT. However, one evaluation component involved using a model from the same family as an evaluator, which introduces a potential bias. When accounting for this limitation, SBERT emerges as a strong alternative, particularly for retrieval-focused tasks. Therefore, while GPT demonstrates superior overall capability, SBERT may be recommended in scenarios where evaluation neutrality and embedding robustness are prioritized.

6.1.3 Layman vs Expert Queries

To investigate how well the system supports laypersons compared to domain experts (RQ3), the results show that GPT and SBERT consistently provide high-quality answers for both layman and expert queries, with semantic similarity and composite scores remaining strong regardless of query type. BioBERT, in contrast, performs significantly worse, particularly in semantic alignment, for both layman and expert questions. Keyword overlap is slightly higher for expert queries, suggesting that expert phrasing aligns better with domain-specific terminology, but overall, the models

maintain comparable performance across layperson and expert questions.

These findings indicate that the system can reliably support layperson queries without substantial loss in content accuracy or semantic relevance, addressing RQ3. Direct query transformation (layman to expert phrasing) was not applied here, but the results suggest that GPT and SBERT are robust enough to handle terminology variations, partially addressing RQ3.2.

6.2 Limitations

This thesis provides an evaluation of selected embedding and language models for biomedical literature and dataset retrieval, but several limitations should be noted. Only three models, GPT-based embeddings with a GPT chat model, Sentence-BERT, and BioBERT were analyzed, leaving out numerous other embeddings (e.g., PubMedBERT, ClinicalBERT) that may perform differently. The evaluation of GPT-based models involved a GPT-family model as an evaluator, introducing potential bias that may overestimate its effectiveness, whereas SBERT emerged as a strong alternative for retrieval-focused tasks. While the system's ability to handle layperson versus expert queries was assessed, a full query transformation pipeline was not implemented, and terminology handling was limited to a curated set of query pairs. Dataset retrieval relied mainly on keyword-based matching, limiting exact matches with gold-standard datasets and leaving advanced semantic or hybrid retrieval unexplored. Overall, the results provide insights into the capabilities of the selected models within the defined scope, but generalization beyond the evaluated models, datasets, and query types should be approached with caution.

6.3 Future Work

For future work, several directions could further enhance the system's capabilities. First, extending the approach to multimodal data, such as images, audio, or other biosignals, could improve retrieval and interpretation beyond text alone. Additionally, improving dataset retrieval performance could be achieved by enhancing dataset annotations or using more advanced semantic or hybrid retrieval methods. Second, the addition of a trained encoder tailored to the specific biomedical domain may address performance limitations observed with models like BioBERT, enabling better semantic understanding and relevance. Third, exploring retrieval methods that incorporate biosignals directly, rather than relying solely on text queries, could provide a more integrated and effective search experience. Finally, while the system currently visualizes datasets, further work on signal processing and interpretation could improve the practical utility of the retrieved data and naturally complement multimodal integration, enabling users to derive more meaningful insights from complex biomedical signals.

References

[1] S. Dash, S. Shakyawar, M. Sharma, et al., "Big data in healthcare: Management, analysis and future prospects", *Journal of Big Data*, vol. 6, no. 54, pp. 1–25, 2019. DOI: 10.1186/s40537-019-0217-0. [Online]. Available: https://doi.org/10.1186/s40537-019-0217-0.

[2] S. P. Bhavnani and A. M. Sitapati, "Virtual care 2.0—a vision for the future of data-driven technology-enabled healthcare", *Current Treatment Options in Cardiovascular Medicine*, vol. 21, no. 5, pp. 1–13, 2019. DOI: 10.1007/s11936-019-0727-2.

[3] T. Sakurai, "Challenges in managing and utilizing biomedical signal data", *NII Journal*, no. 2, pp. 3–12, 2002. [Online]. Available: https://www.nii.ac.jp/journal/pdf/02/02-02.pdf.

[4] A. M. Sitapati et al., "Integrated precision medicine: The role of electronic health records in delivering personalized treatment", *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 9, no. 6, e1378, 2017. DOI: 10.1002/wsbm.1378. [Online]. Available: https://doi.org/10.1002/wsbm.1378.

[5] J. Roh, M. Song, Y. Park, and J. Lee, "Big data and deep learning in biomedical sciences", *Annual Review of Biomedical Data Science*, vol. 3, pp. 1–21, 2020. DOI: 10.1146/annurev-biodatasci-080917-013343. [Online]. Avail-

able: https://www.annualreviews.org/content/10.1146/annurev-biodatasci-080917-013343.

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks", *Information and Medicine Unlocked*, vol. 1, pp. 12–22, 2017. DOI: 10.1016/j.imu.2017.04.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914817300242.

[7] D. H. R. Blackwood and W. J. Muir, "Cognitive brain potentials and their application", *British Journal of Psychiatry*, vol. 157, no. S9, pp. 96–101, 1990. DOI: 10.1192/S0007125000291897.

[8] M. A. Almarshad, M. S. Islam, S. Al-Ahmadi, and A. S. BaHammam, "Diagnostic features and potential applications of ppg signal in healthcare: A systematic review", *Healthcare*, vol. 10, no. 3, p. 547, 2022. DOI: 10.3390/healthcare10030547. [Online]. Available: https://doi.org/10.3390/healthcare10030547.

[9] R. F. Gillum, "From papyrus to the electronic tablet: A brief history of the clinical medical record with lessons for the digital age", *The American Journal of Medicine*, vol. 126, no. 10, pp. 853–857, 2013, ISSN: 0002-9343. DOI: 10.1016/j.amjmed.2013.03.024. [Online]. Available: https://doi.org/10.1016/j.amjmed.2013.03.024.

[10] E. Kim, S. M. Rubinstein, K. T. Nead, A. P. Wojcieszynski, P. E. Gabriel, and J. L. Warner, "The evolving use of electronic health records (ehr) for research", *Seminars in Radiation Oncology*, vol. 29, no. 4, pp. 354–361, 2019, Big Data in Radiation Oncology, ISSN: 1053-4296. DOI: https://doi.org/10.1016/j.semradonc.2019.05.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053429619300426.

[11] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review", *Journal of Healthcare Engineering*, vol. 2018, no. 1, p. 4 302 425, 2018. DOI: https://doi.org/10.1155/2018/4302425. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2018/4302425. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2018/4302425.

[12] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine", *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019. DOI: 10.1056/NEJMra1814259. [Online]. Available: https://doi.org/10.1056/NEJMra1814259.

[13] C. S. Kruse, C. Kristof, B. Jones, E. Mitchell, and A. Martinez, "Barriers to electronic health record adoption: A systematic literature review", *Journal of Medical Systems*, vol. 40, no. 12, p. 252, 2016. DOI: 10.1007/s10916-016-0628-9. [Online]. Available: https://doi.org/10.1007/s10916-016-0628-9.

[14] R. C. Barrows Jr, M. Busuioc, and C. Friedman, "Limited parsing of notational text visit notes: Ad-hoc vs. nlp approaches", in *Proceedings of the AMIA Symposium*, PMID: 11079843; PMCID: PMC2243829, 2000, pp. 51–55. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC2243829/.

[15] E. Ford et al., "Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: What information is hidden in free text?", *BMC Medical Research Methodology*, vol. 13, no. 1, p. 105, 2013. DOI: 10.1186/1471-2288-13-105. [Online]. Available: https://doi.org/10.1186/1471-2288-13-105.

[16] T. Edinger, A. M. Cohen, S. Bedrick, K. Ambert, and W. Hersh, "Barriers to retrieving patient information from electronic health record data: Failure analysis from the trec medical records track", *AMIA Annual Symposium Proceedings*, vol. 2012, pp. 180–188, 2012.

[17] D. A. Hanauer, Q. Mei, J. Law, R. Khanna, and K. Zheng, "Supporting information retrieval from electronic health records: A report of university of michigan's nine-year experience in developing and using the electronic medical record search engine (emerse)", *Journal of Biomedical Informatics*, vol. 55, pp. 290–300, 2015, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2015.05.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046415000829.

[18] S. Kogan, Q. Zeng, N. Ash, and R. A. Greenes, "Problems and challenges in patient information retrieval: A descriptive study", in *Proceedings of the AMIA Symposium*, American Medical Informatics Association, 2001, pp. 329–333. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11825205/.

[19] Q. Zeng, S. Kogan, N. Ash, and R. A. Greenes, "Patient and clinician vocabulary: How different are they?", *Studies in Health Technology and Informatics*, vol. 84, no. Pt 1, pp. 399–403, 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11604772/.

[20] L. Tamine and L. Goeuriot, "Semantic information retrieval on medical texts: Research challenges, survey, and open issues", *ACM Comput. Surv.*, vol. 54, no. 7, Sep. 2021, ISSN: 0360-0300. DOI: 10.1145/3462476. [Online]. Available: https://doi.org/10.1145/3462476.

[21] R. Islamaj Dogan, G. C. Murray, A. Neveol, and Z. Lu, "Understanding pubmed user search behavior through log analysis", *Database (Oxford)*, vol. 2009, bap018, 2009. DOI: 10.1093/database/bap018.

[22] C.-C. Huang and Z. Lu, "Discovering biomedical semantic relations in pubmed queries for information retrieval and database curation", *Database*, vol. 2016, baw025, Mar. 2016, ISSN: 1758-0463. DOI: 10.1093/database/baw025. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baw025/8223375/baw025.pdf. [Online]. Available: https://doi.org/10.1093/database/baw025.

[23] C. E. Crangle, A. Zbyslaw, J. M. Cherry, and E. L. Hong, "Concept extraction and synonymy management for biomedical information retrieval.", in *TREC*, 2004. [Online]. Available: https://trec.nist.gov/pubs/trec13/papers/converspeech.geo.pdf.

[24] Y. Wang et al., "A comparison of word embeddings for the biomedical natural language processing", *Journal of Biomedical Informatics*, vol. 87, pp. 12–20, 2018, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2018.09.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046418301825.

[25] Q. Jin and et al., "Pubmed and beyond: Biomedical literature search in the age of artificial intelligence", *eBioMedicine*, vol. 100, p. 104 988, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10850402/.

[26] S. Z. Shariff et al., "Retrieving clinical evidence: A comparison of pubmed and google scholar for quick clinical searches", *Journal of Medical Internet Research*, vol. 15, no. 8, e164, 2013. DOI: 10.2196/jmir.2624.

[27] X. Zhao, H. Jiang, J. Yin, and et al., "Changing trends in clinical research literature on pubmed database from 1991 to 2020", *European Journal of Medical Research*, vol. 27, p. 95, 2022. DOI: 10.1186/s40001-022-00717-9. [Online]. Available: https://doi.org/10.1186/s40001-022-00717-9.

[28] Z. Lu, "Pubmed and beyond: A survey of web tools for searching biomedical literature", *Database (Oxford)*, 2011. DOI: 10.1093/database/baq036. [Online]. Available: https://doi.org/10.1093/database/baq036.

[29] N. Fiorini, R. Leaman, D. J. Lipman, and et al., "How user intelligence is improving pubmed", *Nature Biotechnology*, 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30272675/.

[30] E. Callaway, D. Cyranoski, S. Mallapaty, and et al., "The coronavirus pandemic in five powerful charts", *Nature*, vol. 579, pp. 482–483, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32203366/.

[31] G. Li, Y. Zhou, J. Ji, and et al., "Surging publications on the covid-19 pandemic", *Clinical Microbiology and Infection*, vol. 27, pp. 484–486, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7506363/.

[32] Q. Chen, A. Allot, and Z. Lu, "Keep up with the latest coronavirus research", *Nature*, vol. 579, p. 193, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32157233/.

[33] Q. Chen, A. Allot, and Z. Lu, "Litcovid: An open database of covid-19 literature", *Nucleic Acids Research*, vol. 49, pp. D1534–D1540, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33166392/.

[34] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: From information retrieval to biological discovery", *Nature Reviews Genetics*, vol. 7, no. 2, pp. 119–129, 2006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16418747/.

[35] J. Ma, X. Wu, and L. Huang, "The use of artificial intelligence in literature search and selection of the pubmed database", *Scientific Programming*, pp. 1–9, 2022. DOI: 10.1155/2022/8855307. [Online]. Available: https://doi.org/10.1155/2022/8855307.

[36] J.-Y. Jung, T. F. DeLuca, T. H. Nelson, and D. P. Wall, "A literature search tool for intelligent extraction of disease-associated genes", *Journal of the American Medical Informatics Association*, vol. 21, no. 3, pp. 399–405, 2014. DOI: 10.1136/amiajnl-2012-001563. [Online]. Available: https://doi.org/10.1136/amiajnl-2012-001563.

[37] Z. Lu, W. Kim, and W. J. Wilbur, "Evaluating relevance ranking strategies for medline retrieval", *Journal of the American Medical Informatics Association (JAMIA)*, vol. 16, no. 1, pp. 32–36, 2009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18952932/.

[38] A. G. Jácome, F. Fdez-Riverola, and A. Lourenço, "Biomedical search engine framework: Lightweight and customized implementation of domain-specific biomedical search engines", *Computer Methods and Programs in Biomedicine*, vol. 131, pp. 63–77, 2016, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2016.03.030. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260715300560.

[39] R. van Hoorn et al., "The development of pubmed search strategies for patient preferences for treatment outcomes", *BMC Medical Research Methodology*, vol. 16, no. 1, p. 88, 2016, ISSN: 1471-2288. DOI: 10.1186/s12874-016-0192-5. [Online]. Available: https://doi.org/10.1186/s12874-016-0192-5.

[40] S. Corrao, D. Colomba, C. Argano, and et al., "Optimized search strategy for detecting scientifically strong studies on treatment through pubmed", *Internal and Emergency Medicine*, vol. 7, pp. 283–287, Jun. 2012. DOI: 10.1007/s11739-012-0773-1. [Online]. Available: https://doi.org/10.1007/s11739-012-0773-1.

[41] K. A. Hambarde and H. Proença, "Information retrieval: Recent advances and beyond", *IEEE Access*, vol. 11, pp. 76 581–76 604, 2023. DOI: 10.1109/ACCESS.2023.3295776.

[42] X. Wang, "The application of nlp in information retrieval", *Applied and Computational Engineering*, vol. 42, pp. 290–297, Feb. 2024. DOI: 10.54254/2755-2721/42/20230795.

[43] J. Ramos, "Using tf-idf to determine word relevance in document queries", in *Proceedings of the First Instructional Conference on Machine Learning*, vol. 242, 2003, pp. 29–48. [Online]. Available: https://www.researchgate.net/publication/228818851_Using_TF-IDF_to_determine_word_relevance_in_document_queries.

[44] P. Cichosz, "Bag of words and embedding text representation methods for medical article classification", *International Journal of Applied Mathematics and Computer Science*, vol. 33, Jan. 2023. DOI: 10.34768/amcs-2023-0043.

[45] S. Sivarajkumar, H. A. Mohammad, D. Oniani, et al., "Clinical information retrieval: A literature review", *Journal of Healthcare Informatics Research*, vol. 8, pp. 313–352, 2024. DOI: 10.1007/s41666-024-00159-4. [Online]. Available: https://doi.org/10.1007/s41666-024-00159-4.

[46] S. Frihat, A. Papenmeier, and N. Fuhr, "Enhancing biomedical literature retrieval with level of evidence and bio-concepts: A comparative user study", in *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*. New York, NY, USA: Association for Computing Machinery, 2025, ISBN: 9798400710933. [Online]. Available: https://doi.org/10.1145/3677389.3702558.

[47] National Library of Medicine (NLM), *Medline*, https://www.nlm.nih.gov/medline/medline_overview.html.

[48] P. Shelke, C. Shewale, R. Mirajkar, S. Dedgaonkar, P. Wawage, and R. Pawar, "A systematic and comparative analysis of semantic search algorithms", *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, pp. 222–229, Oct. 2023. DOI: 10.17762/ijritcc.v11i11s.8094.

[49] U. Naseem, K. Musial, P. Eklund, and M. Prasad, "Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding", in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9206808.

[50] D. Vithanage, P. Yu, L. Wang, et al., "Contextual word embedding for biomedical knowledge extraction: A rapid review and case study", *J. Healthc. Inform. Res.*, vol. 8, pp. 158–179, 2024. DOI: 10.1007/s41666-023-00157-y. [Online]. Available: https://doi.org/10.1007/s41666-023-00157-y.

[51] R. Lebret, "Word embeddings for natural language processing", 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:63947950.

[52] J. X. Morris and A. M. Rush, *Contextual document embeddings*, 2024. arXiv: 2410.02525 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2410.02525.

[53] Q. Liu, M. J. Kusner, and P. Blunsom, *A survey on contextual embeddings*, 2020. arXiv: 2003.07278 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2003.07278.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018. DOI: 10.48550/arXiv.1810.04805. [Online]. Available: https://doi.org/10.48550/arXiv.1810.04805.

[55] A. Miaschi and F. Dell'Orletta, "Contextual and non-contextual word embeddings: An in-depth linguistic investigation", in *Proceedings of the 5th Workshop on Representation Learning for NLP*, 2020, pp. 110–119. DOI: 10.18653/v1/2020.repl4nlp-1.15. [Online]. Available: https://doi.org/10.18653/v1/2020.repl4nlp-1.15.

[56] S. Wang and R. Koopman, "Semantic embedding for information retrieval", in *Proceedings of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR 2017)*, P. Mayr, I. Frommholz, and G. Cabanac, Eds., ser. CEUR Workshop Proceedings, vol. 1823, CEUR-WS, 2017, pp. 122–132. [Online]. Available: http://ceur-ws.org/Vol-1823/.

[57] I. Budiman, D. T. Nugrahadi, M. R. Faisal, and M. Rusli, "A study on effect of generated features from word2vec vectors for text classification", [Online]. Available: https://www.researchgate.net/publication/348404518_A_Study_on_Effect_of_Generated_Features_From_Word2Vec_Vectors_For_Text_Classification.

[58] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing Management*, vol. 24, no. 5, pp. 513–523, 1988, ISSN: 0306-4573. DOI: https://doi.org/10.1016/0306-4573(88)90021-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0306457388900210.

[59] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3.", Jan. 1994. [Online]. Available: https://www.researchgate.net/publication/221037764_Okapi_at_TREC-3.

[60] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau, "Okapi at trec", Special Publication 500-207, Tech. Rep., 1992, pp. 21–30.

[61] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

[62] M. Luo, A. Mitra, T. Gokhale, and C. Baral, "Improving biomedical information retrieval with neural retrievers", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 038–11 046, Jun. 2022. DOI: 10.1609/aaai.v36i10.21352. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21352.

[63] A. Faro, D. Giordano, and C. Spampinato, "Combining literature text mining with microarray data: Advances for system biology modeling", *Briefings in Bioinformatics*, vol. 13, no. 1, pp. 61–82, Jun. 2011, ISSN: 1467-5463. DOI: 10.1093/bib/bbr018. eprint: https://academic.oup.com/bib/article-pdf/13/1/61/584822/bbr018.pdf. [Online]. Available: https://doi.org/10.1093/bib/bbr018.

[64] C. Wu, J. M. Schwartz, G. Brabant, and et al., "Constructing a molecular interaction network for thyroid cancer via large-scale text mining of gene and pathway events", *BMC Systems Biology*, vol. 9, no. Suppl 6, S5, 2015. DOI: 10.1186/1752-0509-9-S6-S5. [Online]. Available: https://doi.org/10.1186/1752-0509-9-S6-S5.

[65] T. Alves, R. Rodrigues, H. Costa, and M. Rocha, "Development of an information retrieval tool for biomedical patents", *Computer Methods and Programs in Biomedicine*, vol. 159, pp. 125–134, 2018, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2018.03.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260717310568.

[66] F. He et al., "Applications of cutting-edge artificial intelligence technologies in biomedical literature and document mining", *Medical Review*, vol. 3, no. 3,

pp. 200–204, 2023. DOI: 10.1515/mr-2023-0011. [Online]. Available: https://doi.org/10.1515/mr-2023-0011.

[67] F. Liu, J. Chen, A. Jagannatha, and H. Yu, "Learning for biomedical information extraction: Methodological review of recent advances", Jun. 2016. DOI: 10.48550/arXiv.1606.07993.

[68] S. Madan, M. Lentzen, J. Brandt, et al., "Transformer models in biomedicine", *BMC Med. Inform. Decis. Mak.*, vol. 24, 2024. DOI: 10.1186/s12911-024-02600-5. [Online]. Available: https://doi.org/10.1186/s12911-024-02600-5.

[69] A. Shewalkar, D. Nyavanandi, and S. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru", *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, pp. 235–245, Oct. 2019. DOI: 10.2478/jaiscr-2019-0006.

[70] N. Gruber and A. Jockisch, "Are gru cells more specific and lstm cells more sensitive in motive classification of text?", *Frontiers in Artificial Intelligence*, vol. 3, 2020. DOI: 10.3389/frai.2020.00040. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2020.00040/full.

[71] S. Bayat and G. Işık, "Assessing the efficacy of lstm, transformer, and rnn architectures in text summarization", *International Conference on Applied Engineering and Natural Sciences*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260221127.

[72] A. Johnson, T. Pollard, and R. Mark, *Mimic-iii clinical database*, 2015. DOI: 10.13026/C2XW26. [Online]. Available: https://doi.org/10.13026/C2XW26.

[73] A. E. W. Johnson et al., "Mimic-iii, a freely accessible critical care database", *Sci. Data*, vol. 3, 2016. DOI: 10.1038/sdata.2016.35. [Online]. Available: https://doi.org/10.1038/sdata.2016.35.

[74] D. Miller, "Leveraging bert for extractive text summarization on lectures", *arXiv preprint arXiv:1906.04165*, 2019. [Online]. Available: https://arxiv.org/abs/1906.04165.

[75] J. Lee et al., "Biobert: A pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020. DOI: 10.1093/bioinformatics/btz682. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682.

[76] M. Habibi et al., "Deep learning with word embeddings improves biomedical named entity recognition", *Bioinformatics*, vol. 33, pp. i37–i48, 2017, Supplementary issue. DOI: 10.1093/bioinformatics/btx243.

[77] S. Pyysalo et al., "Distributional semantics resources for biomedical text processing", in *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, Tokyo, Japan, 2013, pp. 39–43. [Online]. Available: https://academic.oup.com/bioinformatics/article/33/14/i37/3953940.

[78] N. M. Gardazi, A. Daud, M. K. Malik, et al., "Bert applications in natural language processing: A review", *Artificial Intelligence Review*, vol. 58, 2025. DOI: 10.1007/s10462-025-11162-5. [Online]. Available: https://doi.org/10.1007/s10462-025-11162-5.

[79] S. Shen, X. Liu, H. Sun, and D. Wang, "Biomedical knowledge discovery based on sentence-bert", *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, e362, 2020. DOI: https://doi.org/10.1002/pra2.362.

[80] S. Ghosh, "Semantic coupling of scientific literature using sbert: An enhanced model for systematic literature review", [Online]. Available: https://www.researchgate.net/publication/377308246_Semantic_Coupling_of_Scientific_Literature_using_sBERT_An_Enhanced_Model_for_Systematic_Literature_Review.

[81] Q. Chen et al., "An extensive benchmark study on biomedical text generation and mining with chatgpt", *Bioinformatics*, vol. 39, no. 9, btad557, 2023. DOI: 10.1093/bioinformatics/btad557. [Online]. Available: https://doi.org/10.1093/bioinformatics/btad557.

[82] Y. Han, C. Liu, and P. Wang, "A comprehensive survey on vector database: Storage and retrieval technique, challenge", *arXiv preprint arXiv:2310.11703*, 2023. DOI: https://doi.org/10.48550/arXiv.2310.11703.

[83] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, dense, and attentional representations for text retrieval", *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 329–345, Apr. 2021, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00369. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00369/1924040/tacl_a_00369.pdf. [Online]. Available: https://doi.org/10.1162/tacl_a_00369.

[84] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, Apr. 2020. DOI: 10.1109/TPAMI.2018.2889473. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2889473.

[85] A. Andoni, *Locality-sensitive hashing (lsh) home page*, https://www.mit.edu/~andoni/LSH/.

[86] Y. Tian, Z. Yue, R. Zhang, X. Zhao, B. Zheng, and X. Zhou, "Approximate nearest neighbor search in high dimensional vector databases: Current research and future directions.", *IEEE Data Eng. Bull.*, no. 3, pp. 39–54, 2023. [Online]. Available: https://hdl.handle.net/1783.1/137249.

[87] *Faiss*, https://github.com/facebookresearch/faiss.

[88] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks", in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481a3-Abstract.html.

[89] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks", *arXiv preprint arXiv:2005.11401*, 2021. [Online]. Available: https://arxiv.org/abs/2005.11401.

[90] S. Regmi and C. P. Pun, "Gpt semantic cache: Reducing llm costs and latency via semantic embedding caching", *arXiv preprint arXiv:2411.05276*, 2024. [Online]. Available: https://arxiv.org/abs/2411.05276.

[91] F. Bang, "GPTCache: An open-source semantic cache for LLM applications enabling faster answers and cost savings", in *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, L. Tan, D. Milajevs, G. Chauhan, J. Gwinnup, and E. Rippeth, Eds., Association for Computational Linguistics, Dec. 2023, pp. 212–218. DOI: 10.18653/v1/2023.nlposs-1.24. [Online]. Available: https://aclanthology.org/2023.nlposs-1.24/.

[92] K. Hatalis et al., "Memory matters: The need to improve long-term memory in llm-agents", in *Proceedings of the AAAI Symposium Series*, vol. 2, 2023,

pp. 277–280. [Online]. Available: https://doi.org/10.1609/aaaiss.v2i1.27688.

[93] X. Zhou, Z. Sun, and G. Li, "Db-gpt: Large language model meets database", *Data Science and Engineering*, vol. 9, pp. 1–10, Jan. 2024. DOI: 10.1007/s41019-023-00235-6.

[94] G. Li, X. Zhou, and X. Zhao, "Llm for data management", *Proceedings of the VLDB Endowment*, vol. 17, no. 12, pp. 4213–4216, 2024. [Online]. Available: https://www.vldb.org/pvldb/vol17/p4213-li.pdf.

[95] R. Tang, X. Han, X. Jiang, and X. Hu, "Does synthetic data generation of llms help clinical text mining?", 2023. [Online]. Available: https://arxiv.org/pdf/2303.04360.

[96] A. Albalak et al., "A survey on data selection for language models", *arXiv preprint arXiv:2402.16827*, 2024. [Online]. Available: https://arxiv.org/abs/2402.16827.

[97] M. Fan et al., "Cost-effective in-context learning for entity resolution: A design space exploration", in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, IEEE, 2024, pp. 3696–3709. [Online]. Available: https://arxiv.org/pdf/2312.03987.

[98] X. Zhou et al., "D-bot: Database diagnosis system using large language models", *arXiv preprint arXiv:2312.01454*, 2023. [Online]. Available: https://arxiv.org/abs/2312.01454.

[99] X. Huang et al., "Llmtune: Accelerate database knob tuning with large language models", *arXiv preprint arXiv:2404.11581*, 2024. [Online]. Available: https://arxiv.org/abs/2404.11581.

[100] S. Chang and E. Fosler-Lussier, "How to prompt llms for text-to-sql: A study in zero-shot, single-domain, and cross-domain settings", 2023. [Online]. Available: https://arxiv.org/pdf/2305.11853.

[101] C. Whitehouse, M. Choudhury, and A. F. Aji, "LLM-powered data augmentation for enhanced cross-lingual performance", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Association for Computational Linguistics, Dec. 2023, pp. 671–686. DOI: 10.18653/v1/2023.emnlp-main.44. [Online]. Available: https://aclanthology.org/2023.emnlp-main.44/.

[102] S. Frihat and N. Fuhr, "Integration of biomedical concepts for enhanced medical literature retrieval", *International Journal of Data Science and Analytics*, vol. 20, pp. 4409–4422, 2025. DOI: 10.1007/s41060-025-00724-z. [Online]. Available: https://doi.org/10.1007/s41060-025-00724-z.

[103] W. Abdelkader et al., "Machine learning approaches to retrieve high-quality, clinically relevant evidence from the biomedical literature: Systematic review", *JMIR Medical Informatics*, vol. 9, no. 9, e30401, 2021. DOI: 10.2196/30401. [Online]. Available: https://medinform.jmir.org/2021/9/e30401.

[104] Q. Jin et al., "Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval", *Bioinformatics*, vol. 39, no. 11, btad651, 2023. DOI: 10.1093/bioinformatics/btad651. [Online]. Available: https://doi.org/10.1093/bioinformatics/btad651.

[105] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals", *Circulation*, vol. 101, no. 23, e215–e220, 2000 (June 13), Circulation Electronic

Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[106] E. Sayers. "E-utilities quick start". Updated 2018 Oct 24. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK25500/.

[107] P. Docs, *Text-embedding-ada-002*, https://docs.pinecone.io/models/text-embedding-ada-002, Accessed 2025-11-12, 2025.

[108] Sentence-Transformers. "All-mpnet-base-v2". Accessed 2025-11-12. [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[109] DMIS-Lab. "Biobert-base-cased-v1.1". Accessed 2025-11-12. [Online]. Available: https://huggingface.co/dmis-lab/biobert-base-cased-v1.1.

[110] C. Sakai et al., "Dna damage accumulation impedes cardiac repair after myocardial infarction because of insufficient il-10 expression", *International Heart Journal*, vol. 66, no. 5, pp. 883–894, 2025. DOI: 10.1536/ihj.25-306.

[111] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database", *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, May 2001. [Online]. Available: https://physionet.org/content/mitdb/1.0.0/.

[112] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet", *Biomedizinische Technik*, vol. 40, no. Supplement 1, p. 317, 1995. Accessed: Mar. 5, 2026. [Online]. Available: https://physionet.org/content/ptbdb/1.0.0/.

[113] B. Bai et al., "Early assessment and analysis of high-risk factors of neurodevelopmental impairment in neonates with congenital diaphragmatic hernia", *Frontiers in Pediatrics*, vol. 13, p. 1 632 735, 2025. DOI: 10.3389/fped.2025.1632735.

[114] J. Guttag, "CHB-MIT Scalp EEG Database", *PhysioNet*, Jun. 2010, Version 1.0.0. DOI: 10.13026/C2K01R. [Online]. Available: https://doi.org/10.13026/C2K01R.

[115] C. P. Cheung, R. E. Baker, A. M. Coates, and J. F. Burr, "The acute cardiovascular response to dynamic exercise and recovery following cannabis use", *American Journal of Physiology - Heart and Circulatory Physiology*, vol. 329, no. 6, H1655–H1665, 2025. DOI: 10.1152/ajpheart.00608.2025.

[116] B. Moody, G. Moody, M. Villarroel, G. D. Clifford, and I. Silva, "MIMIC-III Waveform Database", *PhysioNet*, Apr. 2020, Version 1.0. DOI: 10.13026/c2607m. [Online]. Available: https://doi.org/10.13026/c2607m.

[117] B. Moody, S. Hao, B. Gow, T. Pollard, W. Zong, and R. Mark, "MIMIC-IV Waveform Database", *PhysioNet*, Jul. 2022, Version 0.1.0. DOI: 10.13026/a2mw-f949. [Online]. Available: https://doi.org/10.13026/a2mw-f949.

[118] PhysioNet, *Wfdb python toolbox*, https://wfdb.io/.

[119] *Streamlit – a faster way to build and share data apps*, https://streamlit.io/.

[120] D. S. Baim et al., "Survival of patients with severe congestive heart failure treated with oral milrinone", *Journal of the American College of Cardiology*, vol. 7, no. 3, pp. 661–670, Mar. 1986, BIDMC Congestive Heart Failure Database. [Online]. Available: https://www.physionet.org/content/chfdb/1.0.0/.

[121] H. Wan et al., "Efficacy and safety of tongmai jiangtang capsule in the treatment of type 2 diabetes mellitus complicated with coronary heart disease with syndrome of damp-heat obstructing collaterals", *Phytomedicine*, vol. 147, p. 157 234, 2025, Epub 2025 Sep 5. DOI: 10.1016/j.phymed.2025.157234.

[122] J. Ren, T. Feng, X. Wu, et al., "Efficacy and safety of traditional chinese medicine in alleviating symptoms associated with myocardial bridge: A systematic review and meta-analysis", *Frontiers in Pharmacology*, vol. 16, p. 1 619 617, 2025. DOI: 10.3389/fphar.2025.1619617.

[123] Z. Xu et al., "Effectiveness and safety analysis of qifu yixin prescription for the treatment of heart failure with preserved ejection fraction: Study protocol for a randomized, double-blind, placebo-controlled clinical trial", *BMC Complementary Medicine and Therapies*, vol. 25, no. 1, p. 345, 2025. DOI: 10.1186/s12906-025-05106-3.

[124] Y. Liao, D. Lu, and J. Yang, "Changes and monitoring technology of human heart rate and blood oxygen saturation under high-altitude hypoxia", *Frontiers in Physiology*, vol. 16, p. 1 642 777, 2025. DOI: 10.3389/fphys.2025.1642777.