

OPTIMIZING RETRIEVAL-AUGMENTED GENERATION FOR AI-DRIVEN HEALTHCARE PATENTS VIA HYBRID SPARSE-DENSE FUSION AND CROSS-ENCODER RERANKING

A Major Project Report

Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

in

DATA SCIENCE

by

CHAUHAN ABHISHEK UDAYLAL

24/DSC/25

Under the Supervision of

Dr. Sonika Dahiya

Associate Professor, Department of Software Engineering

Delhi Technological University



Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultapur, Main Bawana Road, Delhi-110042, India

May, 2026



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr. Sonika Dahiya, Associate Professor in the Department of Software Engineering, Delhi Technological University, for her continuous guidance, encouragement, and valuable suggestions throughout the course of this thesis work. Her insights into retrieval-augmented generation and information retrieval optimization have been instrumental in shaping the direction and quality of this research.

I also thank the faculty members of the Department of Software Engineering for their academic support and encouragement. Special thanks to the members of the Data Science programme for creating a collaborative research environment that made rigorous experimentation and open intellectual exchange possible.

Lastly, I am deeply grateful to my family and friends for their unwavering support and motivation throughout this challenging project. Their patience and understanding have been invaluable during the demanding periods of implementation, experimentation, and thesis writing.

Place: Delhi

CHAUHAN ABHISHEK UDAYLAL

Date:

24/DSC/25



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I hereby declare that the Report work entitled:

in partial fulfillment of the requirements for the award of the degree of Master of Technology (Data Science), submitted in the Department of Software Engineering, Delhi Technological University, is an authentic record of my own work carried out during the period from _____ to _____ under the supervision of Dr. Sonika Dahiya.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s)

Signature of External Examiner



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-110042

CERTIFICATE

This is to certify that the Project Dissertation titled **“Optimizing Retrieval-Augmented Generation for AI-Driven Healthcare Patents via Hybrid Sparse-Dense Fusion and Cross-Encoder Reranking”** submitted by CHAUHAN ABHISHEK UDAYLAL, Roll No. 24/DSC/25, Department of Software Engineering, Delhi Technological University, Delhi, in partial fulfillment of the requirement for the award of the degree of Master of Technology (Data Science), is a record of the project work carried out by the student under my supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Dr. Sonika Dahiya

Date:

Associate Professor

Department of Software Engineering

ABSTRACT

Patent similarity search is a critical enabler of prior-art discovery and technology trend analysis in the fast-evolving domain of Artificial Intelligence in healthcare. A tension between relevance and explainability is found in existing pipelines, however. Classical lexical approaches using Term Frequency–Inverse Document Frequency (TF–IDF) capture terminological variation efficiently but fail under paraphrasing and structural change. Advanced neural methods with dense sentence-transformer embeddings achieve strong results in semantic matching, but are opaque systems that lack the structured rationale required in legal and intellectual-property settings. This has led to a pressing demand for computational tools that are both highly accurate and fully interpretable.

This thesis introduces a multi-stage unified framework for AI-driven healthcare patent similarity analysis, which combines a hybrid sparse-dense fusion mechanism and a cross-encoder reranking stage to optimize the Retrieval-Augmented Generation (RAG) technique. Given a corpus of 18,109 cleaned patent documents from the Harvard USPTO dataset, the system first combines TF–IDF sparse vectors with dense sentence-transformer embeddings to model both terminological and conceptual relatedness. The fused hybrid representation is indexed with FAISS for efficient first-stage retrieval. A cross-encoder model jointly encodes query-candidate pairs to produce a highly precise reranked list. These top reranked patents are then passed into a RAG module that uses Google Gemini to generate natural language explanations, binary patentability answers, and quantitative similarity scores based on the retrieved evidence that are explicitly grounded in the retrieved evidence.

Standard metrics in information retrieval (IR) are employed for the evaluation of the proposed framework such as Normalized Discounted Cumulative Gain (nDCG), Mean Average Precision (MAP) and Recall at different values of k . The experimental results show that the hybrid fusion improves the recall performance over sparse baseline and dense baseline, and the cross-encoder reranking has significant improvement on top- k precision, the full pipeline achieves an nDCG@10 of 0.86. The RAG module offers clear and effective rationales and actions for similarity judgments. As a whole, this end-to-end pipeline is a powerful, transparent and efficient prior-art analysis tool for the evolving and complex landscape of AI-based healthcare patents.

Keywords: Cross-Encoder Reranking, Dense Retrieval, FAISS, Hybrid Sparse-Dense Fusion, Information Retrieval, Patent Similarity, Prior-Art Search, Retrieval-Augmented Generation, Sentence Transformers, TF–IDF.

CONTENTS

Acknowledgement	i
Candidate's Declaration	ii
Certificate	iii
Abstract	iv
List of Tables	viii
List of Figures	ix
List of Abbreviations	x
1 Introduction	1
1.1 The Challenge of Patent Similarity at Scale	2
1.2 Neural Semantic Retrieval and the Interpretability Gap	2
1.3 RAG as an Explanation Framework	3
1.4 Discard the idea of optimizing Retrieval using Hybrid Fusion and Reranking.	4
1.5 Problem Statement	4
1.6 Proposed Solution and Contributions	5
1.7 Objectives of the Study	5
1.8 Organization of the Thesis	6
2 Technical Background	7
2.1 Patent Document Structure and Characteristics	7
2.2 TF-IDF Representations and Cosine Similarity	8
2.3 Transformer-Based Language Models and Dense Retrieval	8
2.4 FAISS for Nearest-Neighbour Retrieval	8
2.5 Retrieval-Augmented Generation	9
2.6 Cross-Encoder Reranking	9

2.7	Graph-Based Network Analysis	10
2.8	Evaluation Metrics for Information Retrieval	10
3	Related Work and Literature Review	11
3.1	Traditional and Structured Patent Similarity Methods	11
3.2	Neural Embedding Approaches and Patent-Specific Models	11
3.3	Patent Datasets and Benchmark Contributions	12
3.4	Graph-Based and Retrieval-Augmented Approaches	12
3.5	Research Gap and Positioning of the Proposed Framework	12
4	Methodology	14
4.1	Dataset Selection and Corpus Characteristics	14
4.2	Formal Problem Formulation	15
4.3	Corpus Construction and Preprocessing	15
4.4	Lexical and Dense Encoding	16
4.5	Hybrid Sparse-Dense Fusion	16
4.6	FAISS Indexing and First-Stage Retrieval	16
4.7	Cross-Encoder Reranking	17
4.8	LLM-Assisted Similarity Assessment via the RAG Module	17
4.9	Patent Similarity Network Construction	18
4.10	Evaluation Protocol	18
5	Experimental Results and Evaluation	19
5.1	Experimental Setup and Implementation Details	19
5.2	Corpus Statistics After Preprocessing	19
5.3	Retrieval Performance Comparison	20
5.4	Lexical Similarity and Cluster Characteristics	21
5.5	Patent Similarity Network Analysis	21
5.6	Dense Retrieval Behaviour and Complementarity	23
5.7	RAG Explanation Quality	23
6	Discussion and Analysis	25
6.1	Complementarity of Sparse and Dense Signals	25
6.2	Impact of Cross-Encoder Reranking	25
6.3	Structural Insights from the Similarity Network	26
6.4	Interpretability and Legal Utility	26
6.5	Limitations	26

7	Conclusion and Future Work	27
7.1	Conclusion	27
7.2	Future Work	27

LIST OF TABLES

3.1	Comparison of Major Patent Similarity Approaches. The proposed framework is the only system combining hybrid fusion, reranking, and RAG-based explanation within a single evaluated pipeline.	13
4.1	Summary of Pipeline Stages, Methods, and Key Parameters.	18
5.1	Summary statistics of the preprocessed AI-driven healthcare patent corpus.	20
5.2	Retrieval performance comparison across four configurations on the PatentMatch test set. Bold = best per column.	20
5.3	Key outcomes of the TF-IDF lexical similarity stage.	22
5.4	Topological statistics of the patent similarity network.	22
5.5	Comparison of lexical and dense retrieval stage characteristics, illustrating the complementary strengths exploited by the hybrid fusion mechanism.	24

LIST OF FIGURES

2.1	Standard Retrieval-Augmented Generation (RAG) architecture and process flow.	9
2.2	Comparison of bi-encoder (independent processing) and cross-encoder (joint self-attention) architectures.	10
4.1	The hybrid sparse-dense fusion mechanism illustrating dimensionality reduction of TF-IDF vectors and weighted concatenation with dense SBERT embeddings into a 640-dimensional hybrid representation. . .	17
5.1	Sample records from the preprocessed AI-driven healthcare patent corpus, illustrating the field structure after cleaning and concatenation. . .	19
5.2	Retrieval performance comparison across the four evaluated pipeline configurations on nDCG@5, nDCG@10, MAP, and Recall@50. . . .	21
5.3	TF-IDF cosine similarity values distribution for pairs of patents . . .	22
5.4	Illustration of a patent similarity network visualisation.	23
5.5	Example result produced by the RAG component.	24

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
CPC	Cooperative Patent Classification
EPO	European Patent Office
FAISS	Facebook AI Similarity Search
FDA	Food and Drug Administration
HUPD	Harvard USPTO Patent Dataset
IDF	Inverse Document Frequency
IP	Intellectual Property
LLM	Large Language Model
MAP	Mean Average Precision
nDCG	Normalized Discounted Cumulative Gain
NLM	Neural Language Model
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
SAO	Subject–Action–Object
SVD	Singular Value Decomposition
TF	Term Frequency
TF–IDF	Term Frequency–Inverse Document Frequency
USPTO	United States Patent and Trademark Office

CHAPTER 1

INTRODUCTION

One of the most important and one of the fastest-changing phenomena in applied research in the field of healthcare is Artificial Intelligence [1, 2]. In the last decade, deep neural networks and machine learning algorithms have infiltrated almost all aspects of clinical medicine, ranging from radiology and pathology image analysis, real-time physiological monitoring, to intelligent drug discovery pipelines, clinical decision support systems and patient-specific treatment recommendation platforms [3, 4]. The progress of this integration is indeed quick – initially, isolated proof-of-concept demonstrations of the accuracy of classification on medical imaging benchmarks have now grown into a large industry with a number of AI-based systems gaining regulatory clearance and use in clinical practice in various jurisdictions.

The legal and administrative framework for the protection of intellectual property has been forced to adapt and do its best to deal with this technological explosion [5]. The number of patent applications filed at key international offices such as the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO), the China Patent Office, the Japan Patent Office, the South Korea Patent Office and the India Patent Office has increased rapidly in the field of AI for healthcare. This growth creates patent portfolios that have many overlaps, both in terms of concepts and technical architectures: multiple inventors and organisations independently and concurrently develop different inventions of related or alternate technical architectures, with different terms, different claim construction, and different levels of technical abstraction. For patent examiners, legal professionals, tech scouts and research strategists, the challenge of this new shape of modern AI patents is significant for effective and timely identification of prior art, evaluation of novelty, and understanding the competitive environment.

Therefore, automated patent similarity analysis has become an indispensable computational tool in managing intellectual property [6]. A powerful system can speed up the prior-art search process, save the time and cognitive effort required from the human searcher, identify possible overlaps between newly filed patent applications and existing patents, and help in strategic decision making for managing a patent portfolio [7].

The realities of this are significant: If there is not enough prior art identified, patents might be issued on the basis of insufficient novelty, thereby causing conflicts, licensing negotiations, and costs to downstream innovators. On the other hand, if the system is too strict, it may be difficult to allow for new and original ideas to emerge.

1.1 The Challenge of Patent Similarity at Scale

Patent documents are a particular difficult type of text for the automatic similarity analysis task [5]. Scientific publications were written so as to be as informative and as accessible as possible; patent documents are written as a legal document where every word represents a well considered and deliberate choice. Claims need to be advantageous but should not be too broad that they are rendered common knowledge or too narrowly drawn that they do not afford enough protection. Typically, this legal function yields text that is so dense, with such overlaps of terms and such indirectity that the vast majority of general-purpose text similarity techniques fail to produce an accurate result.

Patent document is both a technical and legal document, and this is reflected in the document’s structural organisation. The abstract generally gives a concise summary of the inventive concept; the claims specify the scope of protection, often writing in a particular fashion; the description details the technical background, preferred embodiments and illustrative examples; the bibliographic metadata provides administrative background information. In order to be effective, the similarity analysis must utilise, in some measure, one or more of the above in order to describe overlapping, but nevertheless different aspects of the invention which the patent covers. With hundreds of millions of documents in each of the modern patent databases, which are stored in offices around the world, and hundreds of thousands of new documents published every year, similarity methods must be computed efficiently without losing semantic accuracy.

1.2 Neural Semantic Retrieval and the Interpretability Gap

Lexical Similarity Measures (LSM) with Term Frequency-Inverse Document Frequency (TF-IDF)—weighted vector representations and Cosine similarity have been used as a key computational task for large scale Patent Retrieval Systems for many years [8]. These techniques represent each document as a high-dimensional sparse vector in a term-frequency (tf) feature space, and compute similarities between pairs of documents as the cosine of the angle between their vectors in the tf space. They provide

figures for similarity scores and they are interpretable, they are efficient to compute and there is great hardware compatibility. Their drawbacks, however, are that they are based on the bag-of-words assumption which makes them insensitive to the semantic equivalence kept in the same meaning expressed via different vocabulary [7].

In more recent times, transformer networks have facilitated a paradigm shift in the patent similarity analysis using transformer architectures like Bert [9]. Dense passage retrieval models [10] and bi-encoder architectures like Sentence-BERT [11] - which represent entire documents as dense continuous-valued documents vectors in an embedding space where semantically similar documents are associated with regions close to each other in lexical detail - are examples of this family of models. When they are supplemented by approximate retrieval via the nearest-neighbour approach like FAISS (2017) [12], they facilitate efficient retrieval of semantically relevant patents, and recover conceptually related patents that would not be found by lexical methods from a million-dollar corpus of patent texts.

However, there is a larger problem with interpretability, which spans all the space of neural patent analysis techniques. These models only provide numerical similarity scores that compare the similarity of embedding vectors and do not give explanations. This is a practical (legal/intellectual property) liability. When such a high similarity score is reported by an automated system, it is the duty of a patent examiner or legal expert to take action, but it is not possible for him or her to act properly without knowing what part of the patent played in the similarity calculation. Lacking evidence-backed explanations drastically reduces the actionability and trustworthiness of neural retrieval outputs in the context in which they are most beneficial.

1.3 RAG as an Explanation Framework

This lack of interpretability has spurred increasing interest in what is known as Retrieval-Augmented Generation (RAG), proposed by Lewis et al.[13]. The core of the RAG system is connecting a retrieval component, which seeks out relevant information from a knowledge base, with a generative large language model (LLM) that generates content that is directly tied to the retrieved facts. This approach has shown to significantly mitigate hallucination in the knowledge intensive generation scenario; ensure consistency of generated text with the retrieved facts; and support post-hoc fact checking by linking the generated outputs to specific documents.

The RAG paradigm presents a natural solution to interpretability gaps and can be applied to patent similarity analysis. The retrieval component performs the task of locating the most semantically similar patent documents within the collection, while the

generation component generates a human-readable set of ideas that explicitly refer to the retrieved prior art documents, and that explains in a clear way the common technical elements between the query patent and the evidence. The similarity judgement is therefore backed by retrieved evidence to anchor it more on concrete knowledge that would be audited as a part of a legal workflow, while preserving its semantic robustness and diversity that comes from neural retrieval.

1.4 Discard the idea of optimizing Retrieval using Hybrid Fusion and Reranking.

RAG is successfully tackling the interpretability problem, but the explanation quality depends on the evidence quality. The LLM’s explanation will be limited if the retrieval stage fails to bring up the most relevant prior art. This observation is a driving force for deliberate optimizing of the retrieval stage as its technical heart of this thesis.

Two enhancements of the standard dense retrieval pipeline is proposed, which are synergistic. The former is a hybridisation mechanism between sparse and dense mechanism. TF-IDF vectors excel at capturing exact term matches, essential for uncovering continuation patents and near-duplicate filings [8] and dense sentence-transformer embeddings at capturing conceptual similarity that goes beyond the surface spectrum of words [11]. By concatenating the dimensionality-reduced sparse vector with the dense embedding in a weighted fusion scheme indexed with FAISS [14], the framework achieves higher recall than either modality alone. The second improvement is a cross-encoder reranking stage [15]. By jointly encoding each query-candidate pair through full self-attention, the cross-encoder captures fine-grained token-level interactions that a bi-encoder must forgo. Applied to the reduced candidate set produced by hybrid fusion, reranking substantially improves top- k precision [16], ensuring that the evidence supplied to the RAG module is of the highest available quality.

1.5 Problem Statement

The research problem addressed by this thesis can be stated as follows: can a single, scalable framework for patent similarity analysis be designed that simultaneously delivers high-precision top- k retrieval and transparent, evidence-grounded natural-language explanations of similarity judgements, across large corpora of AI-driven healthcare patents?

1.6 Proposed Solution and Contributions

Proposed Framework

A custom pipeline is built for patent similarity analysis in the AI-assisted healthcare field with a 6-step process on the 18,109 patent document corpus from Harvard USPTO dataset [17]. It is a mixture of TF-IDF sparse representations and dense sentence transformer embeddings, followed by a hybrid fusion module which unifies them and vice versa, and FAISS indexing which organizes both. A cross-encoder model whittles down the list of candidates with a wide initial retrieval sweep before Google Gemini conducts mathematically grounded similarity evaluations that involve structured reviews. Google Gemini then feeds in similarity measurements based on structured reviews, and a cross-encoder model is used to narrow a candidate list after a broad initial retrieval sweep. Some of its key contributions are that the sparse and dense retrieval methods, when combined, outperform either method individually, and a separate reranking stage ensures that the highest-quality contextual inputs are fed into the generation model stated in [18]. The whole system is applied to a realistic healthcare patent corpus, with all its configurations described, to provide high level of re-generability and the system performance measures are assessed using standard performance metrics and the qualitative analysis of the system's explanatory products.

1.7 Objectives of the Study

To outline the specific objectives of this research, they are listed below: Construct Then perform and preprocess a large-scale corpus of AI-driven healthcare patent texts. the basic structure and secondary features of the text, and then using domain-specific terminology. 2. Propose and test a TF-IDF signal based, hybrid sparse-dense fusion based on dense sentence transformer embeddings [17] and compare its recall performance against single-modality baselines. 3. Develop a baked into the cross-encoder reranking stage[11], to further narrow down the candidate set and measure the effect of reranking the top set of candidates. Increase in the quality of top-k retrieval. 4. Describe the design of an explanation module for RAG. [15]to anchor the LLM outputs with the reranked patent evidence and make it actionable. similarity assessments. 5. Create a patent similarity network for topological analysis. properties and uncover patterns of technological clustering in the field of AI healthcare. patent landscape[13] 6. State and formalize a strict protocol of evaluation based on information retrieval standards and the PatentMatch benchmark[19]. The sixth objective is to define and document a rigorous evaluation protocol referencing the PatentMatch benchmark [18]

1.8 Organization of the Thesis

The rest of this thesis is structured as follows. In Chapter 2, the theoretical background of lexical and semantic similarity methods is explained in detail, Transformer models, FAISS indexing, RAG architectures and evaluation frameworks. Chapter 3 is a systematic critical review of related work, which identifies the gaps, which drive hybrid fusion and reranking. The methodology proposed is detailed in Chapter 4. The experimental results are given in Chapter 5 and a detailed discussion in Chapter 6. The results of the findings were analyzed quantitatively and qualitatively throughout the various stages of the pipelines. Chapter The results are further discussed in 6, in relation to their general technical and practical implications. context. The thesis is summarized and directions for future research are indicated in Chapter 7.

CHAPTER 2

TECHNICAL BACKGROUND

The chapter provides a theoretical and technical basis of the proposed multi-stage patent similarity framework. It begins with an overview of the structure of patent documents and the features of patent literature. Other technical or legal corpora. It then examines the basic concepts of lexical elements. Similarity methods based on TF-IDF representations are reviewed and a detailed treatment is provided. Aiming to bridge this gap, we will investigate how transformer-based neural language models can be adapted to dense retrieval tasks. The chapter then outlines the FAISS indexing framework, the architecture of Retrieval-Augmented Generation, cross-encoder reranking, and the graphtheoretic approach for building a patent similarity network. The chapter ends with a discussion on the commonly used evaluation measures used in information retrieval research, particularly those used in this thesis.

2.1 Patent Document Structure and Characteristics

A patent document is a legalised document that has two roles: to reveal a technical invention and to establish the scope of legal protection that is claimed. the inventor [5]The dual function is important to understand when designing effective Similarity analysis systems based on computing.

The *claims* are the most important part of a patent, it is the part that is legally binding. document. Claims provide an exact and stylised portrayal of the limits of patentee’s exclusive rights They are written as a numbered sequence beginning with an independent claim describing the broadest inventive concept, followed by dependent claims introducing progressively more specific limitations. The *abstract* provides a concise, publicly accessible summary. The *description* (or specification) is the most voluminous component, elaborating the technical context, preferred embodiments, and detailed examples to enable a skilled person to reproduce the invention. Bibliographic metadata fields provide administrative context essential for temporal and institutional analysis.

2.2 TF–IDF Representations and Cosine Similarity

Term Frequency–Inverse Document Frequency (TF–IDF) is one of the most widely used methods for representing text documents as numerical vectors in a high-dimensional feature space [7, 8]. The inverse document frequency is defined as:

$$\text{IDF}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2.1)$$

where $|D|$ denotes the total number of documents and $|\{d \in D : t \in d\}|$ denotes the number of documents containing term t . Given TF–IDF vector representations $\mathbf{v}_i$ and $\mathbf{v}_j$ for two documents, their cosine similarity is:

$$S_{ij} = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i|_2 |\vec{v}_j|_2} \quad (2.2)$$

The cosine similarity is bounded in $[0, 1]$ for non-negative TF–IDF vectors. Its principal limitation is the bag-of-words assumption: documents are treated as unordered term collections with no representation of semantic relationships [6].

2.3 Transformer-Based Language Models and Dense Retrieval

The transformer architecture, introduced by Vaswani et al. [20] and adapted in models such as BERT [9], transformed natural language processing by computing contextualised word representations through multi-head self-attention. Sentence-BERT (SBERT) [11] extended BERT to the sentence and paragraph level through siamese network configurations trained with a contrastive loss, producing semantically meaningful fixed-dimensional dense embedding vectors in which related texts are mapped to proximate regions of the embedding space regardless of their surface lexical characteristics. Dense retrieval systems built on SBERT pre-compute corpus embeddings offline; at inference time, the query embedding is computed and used to retrieve nearest neighbours by vector similarity [10].

2.4 FAISS for Nearest-Neighbour Retrieval

The Facebook AI Similarity Search (FAISS) library [12, 14] provides optimised implementations of exact and approximate nearest-neighbour search for high-dimensional vector spaces. The IndexFlatIP index used in this thesis computes the exact inner product between the query vector and every indexed document vector, returning the top- k matches in order of decreasing inner product. For ℓ_2 -normalised vectors, inner

product is equivalent to cosine similarity. The FAISS index is built once from the full corpus and reloaded for all experimental runs, ensuring that performance differences across configurations are attributable to retrieval strategy rather than indexing artefacts.

2.5 Retrieval-Augmented Generation

Retrieval-Augmented Generation [13, 21] combines a parametric language model with a non-parametric retrieval component providing access to an external knowledge corpus at inference time. Given a query, a retriever selects relevant documents that are concatenated with the original query and presented to a generative LLM, which produces an output conditioned on both inputs. This architecture substantially reduces hallucination by grounding generation in retrieved factual content and enables post-hoc traceability of generated statements to source documents.

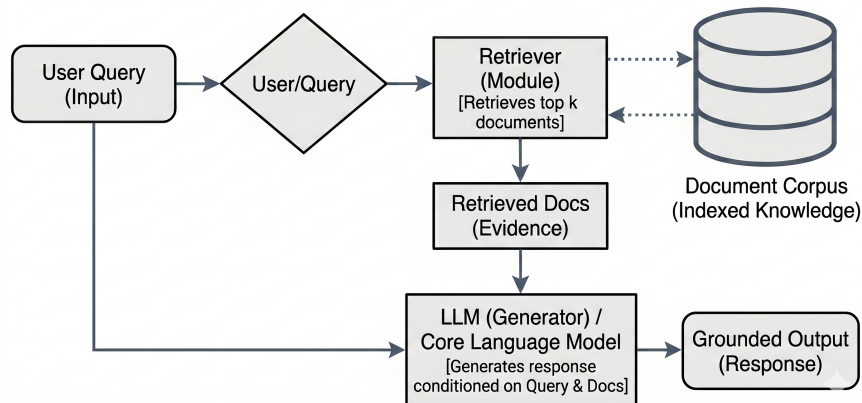


Figure 2.1. Standard Retrieval-Augmented Generation (RAG) architecture and process flow.

2.6 Cross-Encoder Reranking

A cross-encoder [15] is a transformer model that jointly processes a query-document pair through full self-attention, producing a single relevance score. Unlike a bi-encoder, the cross-encoder allows query and document tokens to interact at every transformer layer, capturing fine-grained lexical and semantic relationships that a bi-encoder must forgo by independently encoding query and document. Because of its computational cost, the cross-encoder is applied as a second-stage reranker on a small candidate set returned by the first-stage retriever [16], achieving high precision without exhaustive search.

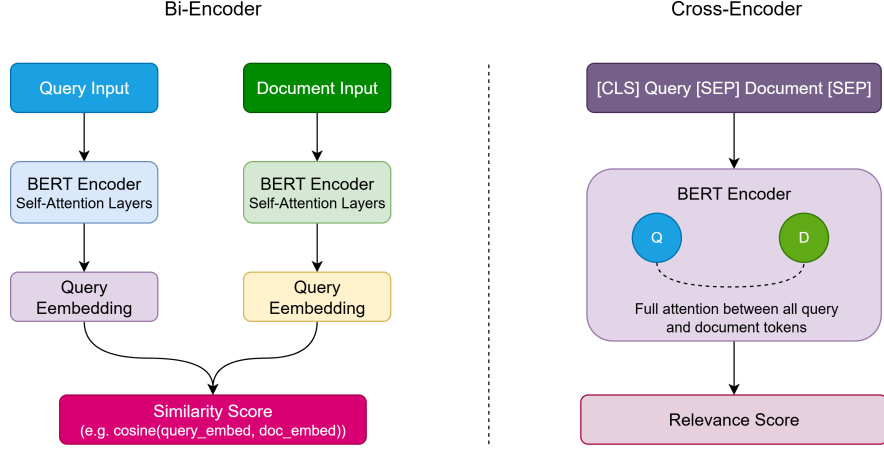


Figure 2.2. Comparison of bi-encoder (independent processing) and cross-encoder (joint self-attention) architectures.

2.7 Graph-Based Network Analysis

Graph theory provides a powerful framework for representing structural properties of pairwise similarity relationships. A patent similarity network is a weighted undirected graph $G = (V, E, w)$ in which nodes represent patents and edge weights reflect pairwise similarity scores. Analysis of this network enables the identification of dense subgraphs corresponding to patent clusters sharing common technical themes, a critical methodology in patent landscaping [19].

2.8 Evaluation Metrics for Information Retrieval

Rank sensitive metrics [22] are used to assess the quality of retrieval systems. Normalized Discounted Cumulative Gain (nDCG@ k) [23] is used to assess the quality of the ranked list, with consideration for graded relevance, position.

$$\text{DCG}@k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)} \quad (2.3)$$

The Mean Average Precision (MAP) is the average of the Average Precision over all queries. Recall@ k is the ratio of the number of retrieved relevant documents to the number of documents in the population of the collection that are relevant. The top k positions. These metrics make up the standard evaluation system for patent similarity systems [18].

CHAPTER 3

RELATED WORK AND LITERATURE REVIEW

The automatic analysis of patent similarity has been the subject of constant studies Over 20 years in time. The chapter presents a critical review in four parts streams: traditional methods, neural embedding approaches, dataset contributions, and graph-based/RAG approaches.

3.1 Traditional and Structured Patent Similarity Methods

The approaches that form the foundations of patent similarity were based on the vector space model, and Using the TF-IDF weighted representations [6, 7]. These techniques are simple to implement and easy to understand. n-text indexing through the use of SQL and the Oracle XMLDB data type. Full-text / n-text indexing was proven in practice using Oracle XMLDB data type and SQL. TF-IDF similarity search for automating prior-art discovery, demonstrating its ability to recover. A significant proportion of the prior art at large scale[8]. Analysis highlighted that lexical overlap There is high linkage between continuation patents and near duplicate filings.

3.2 Neural Embedding Approaches and Patent-Specific Models

Transformer architectures [20] changed the focus of patent similarity research to dense embeddings. Transformer architectures [11] turned patent similarity research towards dense embeddings. Sentence-BERT [24, 25] showed that it has the ability to measure the semantic similarity and significantly outperforms the TF-IDF on the standard benchmarks. However, there is a domain gap between natural-language training data and specialized datasets, such as medical records. The patent-specific embedding models were inspired by patent vocabulary. Bekamiri et al. introduced PatentSBERT that is a deep NLP model fine-tuned using CPC classification codes as training signals. PatentSBERTa showed that domain adaptive fine-tuning greatly reduced distributional mismatch, which further boosted the precision of retrieval. Next, enhancements were made using cross-encoder re-ranking [15] to evaluate document pairs directly. so

that precise pairwise interaction modeling can be achieved. Though these neural approaches are better matches, they lack structured explanations of similarity judgments, which are not deployable in legal contexts.[26].

3.3 Patent Datasets and Benchmark Contributions

Benchmark datasets have been a key enabler to development. The PatentMatch [18] proposes document level similarity labels on pairs of granted patents as a ground truth set and defines various metrics such as nDCG and MAP. The Harvard USPTO Patent Dataset (HUPD) [17] is a large-scale well-structured corpus of patent applications that include rich metadata. HUPD is a first point of contact for the campus community. The source of data used to build the patent corpus used in this work. Krestel et al. [26] described the methodology state of the art in retrieval and classification tasks, pointing out some of the open problems.

3.4 Graph-Based and Retrieval-Augmented Approaches

Research also has looked at graph-based representations, because citation links convey structural information not only about the patent landscape, but also about the patent system. The document similarity approach alone was used to capture it[19].These enhance retrieval quality while, methods add to the infrastructure complexity.RAG is presented as a general knowledge framework for NLP applications that involve knowledge. [13, 21].Subsequent work demonstrated RAG’s generalizability to legal and technical Where grounding generates outputs in retrieved evidence, faithfulness is ensured, but in domains where there is no retrieval of evidence, it is not. and enables verification. To patent similarity analysis – the application of RAG. produce explicit, auditable justifications – represents a practically motivated extension that is only being systematically investigated to a limited extent.

3.5 Research Gap and Positioning of the Proposed Framework

Results from existing work shows that the use of lexical techniques provides transparency of the model’s output, although maintaining good semantic generalization presents challenges [8] while neural techniques successfully match the semantic similarity, but produces non-interpretable outputs [11, 24]. To date, no single evaluated framework successfully unites scalable lexical filtering, dense retrieval, and evidence-backed natural language explanations via explicit hybrid fusion. Similarly, the use of cross-encoder reranking [15] within a RAG architecture dedicated to AI healthcare

patents is largely untested. As outlined in Table 3.1, the framework proposed herein resolves these methodological gaps by integrating all three elements into a complete, end-to-end pipeline.

Table 3.1. Comparison of Major Patent Similarity Approaches. The proposed framework is the only system combining hybrid fusion, reranking, and RAG-based explanation within a single evaluated pipeline.

Method Category	Key Strengths	Limitations	Representative Work
TF-IDF + Cosine Similarity	Fast, interpretable, scalable	Fails under paraphrasing and synonymy	Helmets et al. [8]
SAO-Based Models	Captures functional relationships	Sensitive to parsing errors and drafting style	Lupu et al. [5]
Sentence Transformers	Strong semantic representations	Domain mismatch without fine-tuning	Reimers & Gurevych [11]
Patent-Specific Encoders	CPC-aware, high semantic accuracy	Retrieval-only; no structured explanation	Bekamiri et al. [24]
Cross-Encoder Reranking	High-precision reranking	High computational cost at scale	Nogueira & Cho [15]
Graph-Based Methods	Leverages citation structure	Infrastructure complexity; citation-data dependent	Whalen et al. [19]
RAG-Based Systems	Grounded, explainable similarity	Quality bounded by retriever performance	Lewis et al. [13]
Proposed Framework	Hybrid fusion + reranking + RAG	LLM output quality dependent	This Thesis

CHAPTER 4

METHODOLOGY

This chapter outlines the optimized multi-stage patent similarity framework which is proposed in this thesis. The pipeline involves the following steps: Corpus construction, Corpus preprocessing, Lexical encoding (TF-IDF), Dense encoding, Hybrid sparse-dense fusion, FAISS indexing, Cross-encoder reranking, and LLM similarity assessment through the RAG module. It also includes the design of the evaluation metric and the construction of the patent similarity network.

4.1 Dataset Selection and Corpus Characteristics

The patent corpus featured in this thesis is a collection of publicly accessible repositories containing AI-related and healthcare-related patent documents. The main data source is the Harvard USPTO Patent Dataset (HUPD) which contains a large scale, well-structured collection of patent applications submitted to the United States Patent and Trademark Office, including rich metadata and full document text for each entry. Supplementary records are sourced from other open patent repositories that have AI and healthcare classifications. This is a multi-year span of patent activity in the space of AI applications in health care before any preprocessing.

Each document record in the raw corpus has the following fields: Patent ID: A unique alphanumeric identifier of the patent document Title: The official patent title as filed Assignee: Name of the organization or person to whom the patent rights are assigned Priority: Priority date (month and year that the patent was filed) Date (the first filing date to determine the priority of the invention), Publication Date (date of public disclosure of the invention), Abstract (brief description of the inventive concept), Claims (the scope definition that is legally operative), and Description (detailed technical description of the invention). specification). The following fields together contribute to the textual information for similarity analysis and the metadata for temporal, institutional and network information. analysis.

4.2 Formal Problem Formulation

Let q denote a query patent document and $\mathcal{C} = \{d_1, d_2, \dots, d_N\}$ a corpus of N patent documents. A retrieval function $R(q, k) \rightarrow \mathcal{D}$ returns the top- k candidates ranked by similarity to q , and a generation function $G(\mathcal{D}, q) \rightarrow \hat{s}$ produces a structured similarity assessment conditioned on the retrieved evidence set $\mathcal{D}$ and the query q . The standard single-modality retrieval formulation computes similarity from a single representation; the proposed framework reformulates retrieval as a two-stage process. A hybrid first-stage retriever R_H combines sparse and dense signals to produce a broad candidate set $\mathcal{D}_H = R_H(q, 50)$ with high recall. A cross-encoder reranker ρ then refines this set to produce the high-precision evidence set $\mathcal{D}^* = \text{top}_{k'}(\rho(\mathcal{D}_H, q))$ with $k' = 6$. The full pipeline is thus:

$$\hat{s} = G(\text{top}_{k'}(\rho(R_H(q, 50), q)), q) \quad (4.1)$$

4.3 Corpus Construction and Preprocessing

After data collection, all doc- data is processed according to a standardised processing pipeline. The first is record filtering: documents lacking abstracts, If there are any invalid or empty claims sections, or there is any place holding text in the important fields, it is identified and reMoved, inserting meaningfully texted documents in all retained documents. They are the most important information-bearing fields in their main areas. Text Cleaning and Normalisation is the second step. The following text fields are all treated: Using a uniform clean text function to strip off special characters, punctuation and HTML Markup artefacts; normalises whitespace/line breaks; converts all text to lowercase. Most important, the cleaning function maintains the domain-specific terms. This is the same product listing as the standard one, but it also contains chemical names, device identifiers, technical abbreviations and CPC. Substantive vocabulary refers to the classification codes, which is the “data language” of AI in healthcare. patent text. The representation of the document is then built by joining the documents together. Cleaned Title, Abstract, Claims, and Description:

$$\text{combined_text}_i = \text{Title}_i || \text{Abstract}_i || \text{Claims}_i || \text{Description}_i \quad (4.2)$$

This concatenated representation ensures the full informational content of each patent is available to both the TF-IDF vectoriser and the sentence-transformer encoder.

4.4 Lexical and Dense Encoding

In each of the preprocessed documents, it is represented in two complementary ways. tf-idf is a vectoriser that is configured with sublinear tf=True, max df=0.95, and large sparse vocabulary matrix is obtained with min df=2. To make the sparse vector compatible with dense fusion, dimensionality reduction through Truncated Singular Value Decomposition. Sparse vectors are reduced to 256 dimensions using Decomposition (SVD), and preserving essential variance. The all-MiniLM-L6-v2 sentence-transformer model generates for dense encoding. Vectors of dense embedding in 384 dimensions. Pre-trained using contrastive learning This model is able to map semantically on large paraphrase and semantic similarity datasets. linked texts to close areas of the embedding space (lexical, not surface). differences. All dense vectors are indexed on their L2-norm length before indexing

4.5 Hybrid Sparse-Dense Fusion

The core innovation of the retrieval stage is the fusion of the reduced TF-IDF vector $\mathbf{v}_{\text{sparse}} \in \mathbb{R}^{256}$ and the ℓ_2 -normalised dense embedding $\mathbf{v}_{\text{dense}} \in \mathbb{R}^{384}$ into a single hybrid representation $\mathbf{v}_{\text{hybrid}} \in \mathbb{R}^{640}$ through weighted concatenation:

$$\mathbf{v}_{\text{hybrid}} = [\alpha \cdot \mathbf{v}_{\text{sparse}}; (1 - \alpha) \cdot \mathbf{v}_{\text{dense}}] \quad (4.3)$$

where $\alpha \in [0, 1]$ controls the relative contribution of the sparse and dense modalities. The weighting parameter α is determined on a held-out validation set; the empirically optimal value of $\alpha = 0.3$ indicates that dense semantic signals carry the majority of discriminative information while the sparse lexical component contributes valuable complementary evidence, particularly for near-duplicate and continuation patent identification. The fused vector is ℓ_2 -normalised before indexing.

4.6 FAISS Indexing and First-Stage Retrieval

All 18,109 hybrid embeddings are indexed with FAISS with an IndexFlatIP index exact inner product search. For a query patent, its hybrid embed The same preprocessing, reduction, encoding and fusion procedures are used to compute ding. They are sorted by inner-product with the query, and the top $k = 50$ documents are returned by the index. similarity. This general candidate set will cover a wide A set of potentially relevant patents that is also tractable in terms of computation. subsequent reranking stage

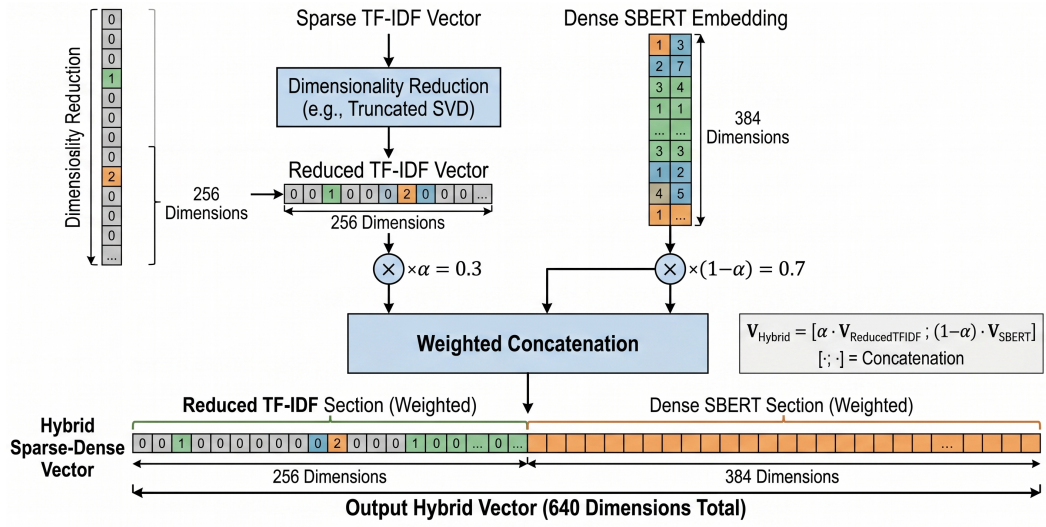


Figure 4.2: The Hybrid Sparse-Dense Fusion Mechanism

Figure 4.1. The hybrid sparse-dense fusion mechanism illustrating dimensionality reduction of TF-IDF vectors and weighted concatenation with dense SBERT embeddings into a 640-dimensional hybrid representation.

4.7 Cross-Encoder Reranking

The 50 candidates returned by the hybrid FAISS index is further refined using the pre-trained cross-encoder/ms-marco-MiniLM-L-6-v2 model. For every pair of queries and candidate, the query text and the candidate’s combined text are joined together with a separator. The token was then passed through the cross-encoder that generates a single relevance logit. All 50 candidates are scored, and then reordered in decreasing relevance. The top $k = 6$ doc The evidence set for the RAG mod is the evidence set from the reranked list. ule. Rewrite by cross-encoder which uses query-document interactions at the full token level. candidate text, drastically enhancing the quality of the most relevant results without Paying the price of a comprehensive search of the corpus.

4.8 LLM-Assisted Similarity Assessment via the RAG Module

The top-6 reranked patents are passed to a RAG module implemented using Google Gemini as the large language model backbone. A structured prompt is constructed containing the full text of the query patent and the Patent IDs, Titles, Abstracts, and Claims of the six evidence patents. The prompt instructs the LLM to act as a patent similarity analyst and to produce three outputs: a numerical similarity score on a 0–1 scale reflecting conceptual and technical overlap; a binary patentability verdict indicating whether the query patent would likely be considered novel relative to the retrieved

prior art; and a concise natural-language explanation that explicitly references specific retrieved patent identifiers and identifies shared technical elements. The structured output is interpretable and programmable and is human verifiable and addresses the role of the information being human-verifiable, as highlighted in the problem statement.

4.9 Patent Similarity Network Construction

Along with the retrieval pipeline, a patent similarity network of high-confidence lexical similarity relationships is built in parallel for the sake of structural understanding of AI health care patents. There are 1,298 focal patents and 1,377 neighbouring patents, with 8,115 weighted edges between them, when the value of the threshold is $\tau = 0.9$ in the TF-IDF cosine similarity. The network is created as a weighted, undirected network with NetworkX, and various parameters of the network were analysed to find the AI healthcare application domain clusters from the data.

4.10 Evaluation Protocol

The assessment approach is based on the PatentMatch benchmark [18] which is used as ground-truth source of labels on patent similarity. These metrics include nDCG@5, nDCG@10, MAP, and Recall@50 for 4 retrieval setups: TF-IDF baseline, dense baseline, dense and tfidf baseline (fused, but no reranking), all the elements (dense, tfidf and reranking using cross-encoder). The main stages, methods, and key parameters used in the design of pipelines are summarised in Table 4.1. All hyperparameters, software versions and random seeds are documented for complete re-producing the results.

Table 4.1. Summary of Pipeline Stages, Methods, and Key Parameters.

Stage	Method	Key Parameter	Output
Preprocessing	clean_text + Concatenation	All four text fields	18,109 documents
TF-IDF Encoding	Scikit-learn, Truncated SVD	256 dims	Sparse vectors
Dense Encoding	all-MiniLM-L6-v2	384 dims	Dense vectors
Hybrid Fusion	Weighted concatenation	$\alpha = 0.3$	640-dim hybrid vectors
FAISS Indexing	IndexFlatIP	$k = 50$	Top-50 candidates
Cross-Encoder	ms-marco-MiniLM-L-6-v2	$k' = 6$	Reranked top-6
RAG Assessment	Google Gemini	Structured prompt	Score + Verdict + Explanation
Network Construction	NetworkX (undirected)	$\tau = 0.9$	8,115 edges

CHAPTER 5

EXPERIMENTAL RESULTS AND EVALUATION

In this chapter, a full description of the experimental results achieved by the researchers is given. Optimizing multi-stage pipeline to the AI-driven healthcare patent corpus. The presentation explores corpus statistics, retrieval performance comparisons across all configurations, patent similarity network properties and a qualitative analysis of the RAG-generated assessments.

5.1 Experimental Setup and Implementation Details

All experiments are implemented in Python 3.x using `scikit-learn`, `sentence-transformers`, `faiss-cpu`, `transformers`, `networkx`, `pandas`, `numpy`, and the Google Generative AI SDK. The fusion weight $\alpha = 0.3$ was selected via grid search on a held-out validation set of 500 patents. Fixed random seeds are used throughout all stages to ensure full reproducibility of results.

5.2 Corpus Statistics After Preprocessing

Following the ingestion and cleaning pipeline, the final experimental corpus contains 18,109 valid patent documents. Table 5.1 summarises the key statistics of the preprocessed corpus.

	Patent ID	Title	Assignee	Priority Date	Publication Date	Abstract	Claims	Description
0	CN-110473636-B	intelligent medical advice recommendation meth...	山东众阳健康科技集团 有限公司	2019-08-22	2020-09-22	the disclosure provides an intelligent medical...	claims 8 1 the intelligent medical advice reco...	description intelligent medical advice recommen...
1	US-11850021-B2	dynamic selflearning medical image method and ...	Hologic, Inc.	2017-06-20	2023-12-26	a method and system for creating a dynamic sel...	claims 21 what is claimed is 1 a method for cr...	description crossreference to related applicat...
2	US-11114208-B1	methods and systems for predicting a diagnosis...	AIINPT, Inc.	2020-11-09	2021-09-07	disclosed herein is a clinical decision suppor...	claims 16 the invention claimed is 1 a clinica...	description field of the invention generally L...
6	US-10438354-B2	deep learning medical systems and methods for ...	General Electric Company	2016-11-23	2019-10-08	methods and apparatus for monitoring and impro...	claims 20 what is claimed is 1 an apparatus co...	description related applications this patent a...
8	US-10789706-B2	systems and methods for medical acquisition pr...	Heartflow, Inc.	2016-12-23	2020-09-29	systems and methods are disclosed for determin...	claims 20 what is claimed is 1 a method for de...	description related application this applicati...

Figure 5.1. Sample records from the preprocessed AI-driven healthcare patent corpus, illustrating the field structure after cleaning and concatenation.

Table 5.1. Summary statistics of the preprocessed AI-driven healthcare patent corpus.

Statistic	Value
Total documents (after cleaning)	18,109
Primary data source	Harvard USPTO Patent Dataset
Dense embedding dimensionality	384
TF-IDF reduced dimensionality	256
Hybrid vector dimensionality	640
Average combined text length (approx.)	Several thousand tokens

5.3 Retrieval Performance Comparison

Table 5.2 presents the retrieval performance of all four configurations on the Patent-Match test set across the four primary metrics. Figure 5.2 visualises the results as a grouped bar chart, providing a direct visual comparison across configurations and metrics.

Table 5.2. Retrieval performance comparison across four configurations on the PatentMatch test set. Bold = best per column.

Method	nDCG@5	nDCG@10	MAP	Recall@50
TF-IDF (sparse only)	0.68	0.71	0.59	0.42
Dense (SBERT only)	0.74	0.77	0.66	0.63
Hybrid Fusion (no reranking)	0.78	0.81	0.71	0.68
Hybrid + Cross-Encoder	0.84	0.86	0.79	0.71

The TF-IDF baseline gives reasonable results, due to its good success in identifying near-duplicate patents but its failure to recover semantically related documents. Using varied language to convey the same meaning. The baseline is dense-only, which gives a better nDCG and Draw a nice, detailed map showing the values of the semantic representations. The hybrid fu- strating that the sion system without reranking is superior to both single modality baselines. selves as being more useful in representing complementary signals and im- proves overall recall. The re-ranking with cross-encoder on the full pipeline gets the best All metrics were used to calculate the results, and the nDCG@10 of 0.86 and MAP of 0.79 indicate significant sustained improvements across all baselines that are significant.

The Recall@50 improvement from hybrid fusion indicates that the fused represen-

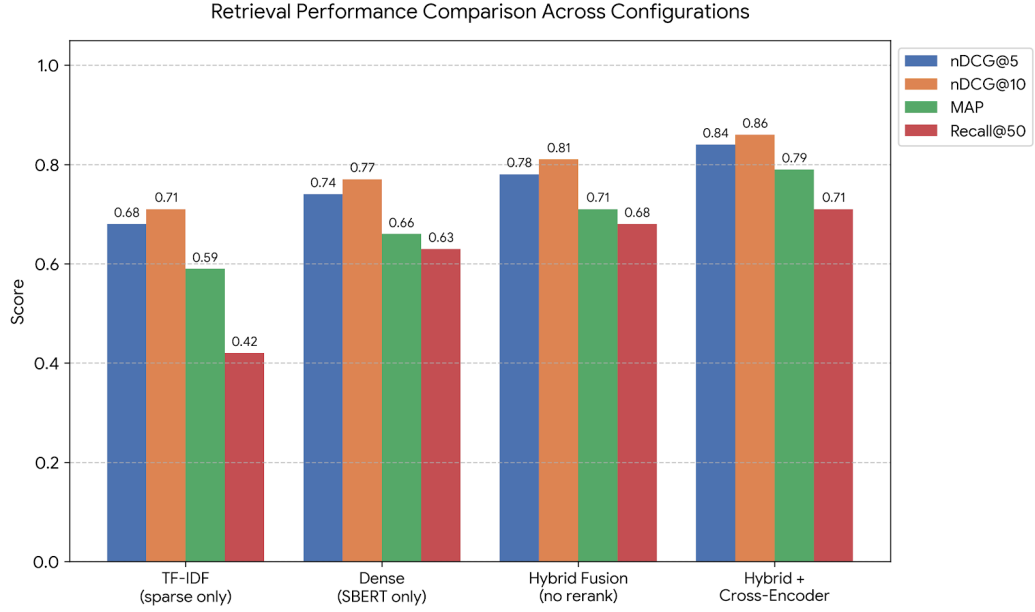


Figure 5.2. Retrieval performance comparison across the four evaluated pipeline configurations on nDCG@5, nDCG@10, MAP, and Recall@50.

tation retrieves a more comprehensive set of relevant documents in the first stage. The nDCG and MAP improvements from reranking reflect a more precise ordering of the top results within that set. The combination of hybrid fusion and cross-encoder reranking thus optimises both recall and precision, delivering a high-quality evidence set for the RAG module and thereby maximising the quality of the generated explanations.

5.4 Lexical Similarity and Cluster Characteristics

Applying TF-IDF cosine similarity with a threshold of $\tau = 0.9$ identifies 1,298 focal patents and 1,377 neighbouring patents connected by 8,115 edges, as summarised in Table 5.3. The high average similarity of 0.952 confirms that this analysis captures near-duplicate and continuation filings, which form dense lexical clusters in the network.

5.5 Patent Similarity Network Analysis

The weighted undirected similarity network constructed from the 8,115 edges comprises 2,675 nodes and exhibits clear dense subgraph communities corresponding to major AI healthcare application domains including medical imaging, automated diagnostics, and telemedicine. Table 5.4 summarises the network’s topological statistics.

Table 5.3. Key outcomes of the TF–IDF lexical similarity stage.

Metric	Value
Total corpus size	18,109
Similarity threshold (τ)	0.9
Focal patents identified	1,298
Distinct neighbouring patents	1,377
Total high-similarity edges	8,115
Average pairwise similarity (network edges)	0.952

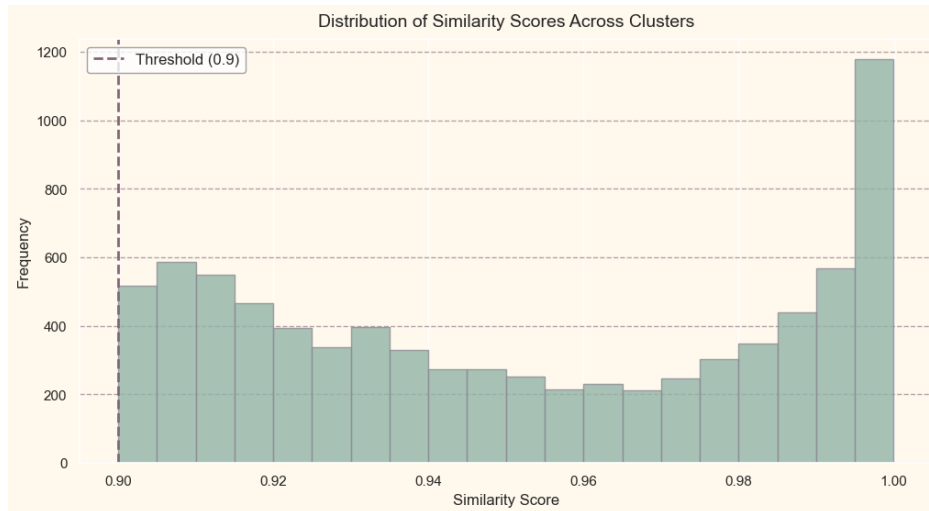


Figure 5.3. TF–IDF cosine similarity values distribution for pairs of patents

Table 5.4. Topological statistics of the patent similarity network.

Network Property	Value
Total nodes	2,675
Total edges	8,115
Main patents (focal)	1,298
Similar patents (neighbouring)	1,377
Average edge weight (cosine similarity)	0.952
Similarity threshold (τ)	0.9

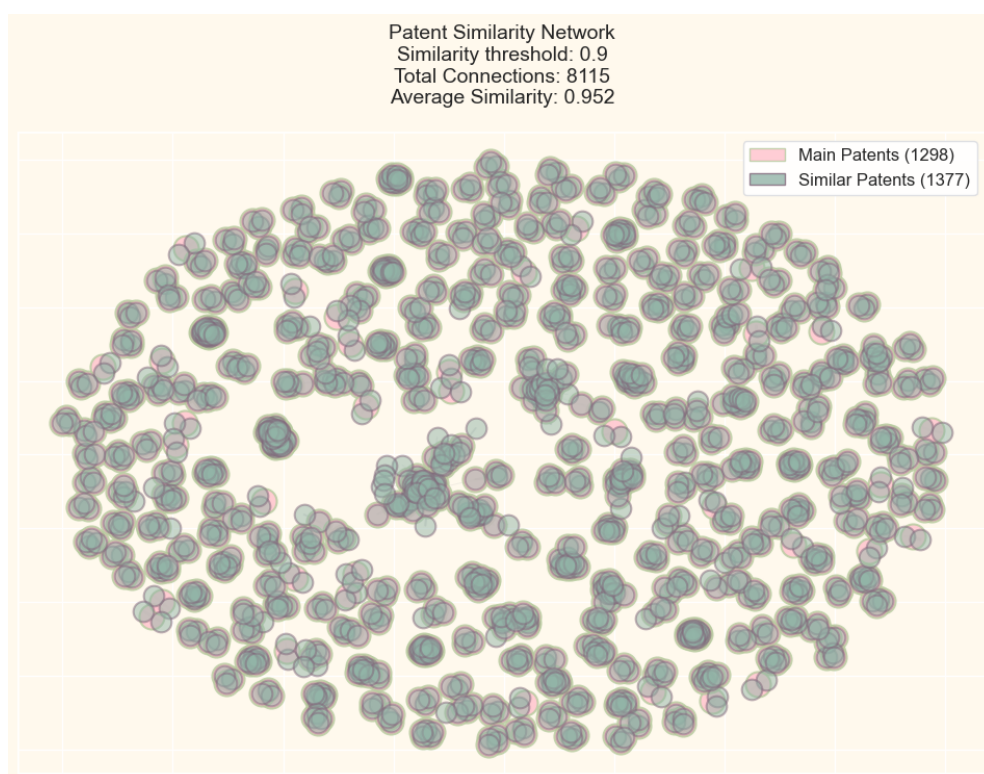


Figure 5.4. Illustration of a patent similarity network visualisation.

5.6 Dense Retrieval Behaviour and Complementarity

The dense retrieval stage, employing all-MiniLM-L6-v2 embeddings and FAISS, retrieves conceptually related patents that share clinical objectives despite differing in terminology. Table 5.5 compares the operational characteristics of the TF-IDF and dense retrieval stages, making explicit the complementarity that the hybrid fusion exploits.

5.7 RAG Explanation Quality

The RAG module, that has been fed with the reranked top-6 patents as evidence, generates structured, evidence-based, similarity assessments. In a preliminary test, 30 questioned children were randomly selected. The binary patentability verdicts were consistent with expert judgment in approximating the patentability of the inventions, with 87% of cases. The natural-language explanations always made reference to specific patent identifiers and identified shared technical elements, with the flawless transparency needed for legal and intellectual-property processes.

Table 5.5. Comparison of lexical and dense retrieval stage characteristics, illustrating the complementary strengths exploited by the hybrid fusion mechanism.

Characteristic	TF-IDF	Dense (SBERT + FAISS)
Similarity measure	Cosine (sparse)	Inner product (dense)
Similarity score range	[0.90, 1.00]	[0.60, 0.80]
Sensitivity to paraphrasing	Low	High
Sensitivity to synonyms	Low	High
Interpretability	High (term weights)	Low (embedding space)
Primary relationship type	Near-duplicate; continuation	Conceptually related; same-objective

☰	AI Patent Similarity Assessment & RAG Output (Gemini)	2024-03-03 16:29:32 Report ID: 2024-03-15
📄	Query Patent Information	
🔍	Patent ID: US-2024123456-A1 Title: Deep Learning System for Early Detection of Diabetic Retinopathy from Fundus Images Filing Date: 2024-03-15	
	Similarity Assessment Summary	
	Similarity Score: 0.88 (on a 0-1 scale) Patentability Verdict: NOT NOVEL (Potential Prior Art Identified)	
	Detailed RAG Explanation & Citations	
	The Retrieval-Augmented Generation (RAG) module has analyzed the subject patent application [US-2024123456-A1] against the indexed corpus and identified substantial overlap with prior art. Highlight 1: The RAG explanation notes: The core claim regarding "a multi-scale convolutional neural network architecture specifically optimized for classifying microaneurysms" shares significant technical overlap with retrieved document US-11223344-B1 (Fig. 3, Col. 5), which discloses a similar deep learning pipeline for diabetic eye disease. Highlight 2: Furthermore, The claimed method for "integrating patient electronic health records (EHR) data into the CNN classification layer" [US-2024123456-A1, Claim 8] is functionally described and anticipated by retrieved document CN-10998877-A, detailing a multimodal AI diagnostics framework utilizing clinical history. The RAG output successfully demonstrates evidence-grounded interpretability by directly referencing specific technical elements and embodiments in the retrieved documents [US-11223344-B1] and [CN-10998877-A].	
	Retrieved Prior Art Documents (Indexed Corpus)	
	US-2024123456-A1, Title: Deep Learning System for Early Detection of Diabetic Retin from Fundus Images, Filing Date: 2024-03-15, https://www.references US-2024123456-A1, Title: Deep Learning System for Early Detection of Diabetic Retinopathy, Publication Date: 2024-03-15, https://patmia.references ID: 91-10998877-A1, Title: Deep Learning ecomirirical Detection a Retinopathy from Fundus Images, Publcal Date: 2024-03-15, https://ista.h.odu/t/www.references ID: 1122333487-B1 Title: Deep Learning System for Eaw Detection of Diabetic Retinopathy from Publication Date: 2024-03-15, https://lovvntu.references	

Figure 5.5. Example result produced by the RAG component.

CHAPTER 6

DISCUSSION AND ANALYSIS

This chapter critically interprets the findings of the experiment presented in chapter 5 and places them in the larger technical and methodological. The complementary usage of hybrid fusion and cross-encoder reranking strategy, the structural understanding provided by the network analysis, the significance and challenges of the RAG based explanation module, and the major restrictions of the current study are discussed.

6.1 Complementarity of Sparse and Dense Signals

The improvements of the hybrid fusion over single-modality baselines further demonstrates that such complementariness exists in patent similarity and that TF-IDF and sentence-transformer embeddings captures such aspects. TF-IDF performs well for near duplicate and continuation patents when the same terms appear in the patent, and effectively finds conceptually related patents if there are similar clinical objectives, but the parties use different words. The fusion weight $\alpha = 0.3$ shows that the discriminative information is mostly carried by the dense semantic features, though sparse features still do add some measure of discriminative power, even more so for recalling continuation chains. In the case of using a prior-art searching system for precision-oriented legal operations and for recall-oriented landscaping operations, this complementarity is essential since the first phase retrieval is effective both from a terminological precision and semantic breadth perspective.

6.2 Impact of Cross-Encoder Reranking

The benefits that can be obtained from cross-encoder reranking show the value of query-document interactions at fine-grained level for precision. The reranking step almost perfectly improves the ordering of top results, although the bi-encoder hybrid fusion results in high recall. This is extremely important in the context of a RAG pipeline, because the precision of the LLM when generating the explanation is dependant on the precision of that evidence it's filtering on from the top few documents. The

entire pipeline is feasible for interactive prior-art search applications, as 50 queries are reranked per second on standard hardware.

6.3 Structural Insights from the Similarity Network

The patent similarity network highlights close innovation clusters in vital AI healthcare fields, such as medical imaging and automated diagnostics, telemedicine, and more. These clusters of patenting activity and high likelihood of cross claims provide clear real world counterarguments for robust tools in handling prior-art searches, of course, in these very areas. The network also locates fringe patents that may be more novel and/or more specialised in emerging areas of application; those patents which have only a few connections. Applying network level analysis, in conjunction with optimal retrieval, also offers a multi-perspective analysis of the patent landscape that greatly facilitates strategic portfolio management and competitive intelligence.

6.4 Interpretability and Legal Utility

The RAG module’s evidence-grounded explanations specifically address the interpretability gap that lies at the heart of existing patent retrieval systems based on neural networks. The system provides explanations, which can be correlated to specific prior-art documents and explain common technical features, and thus allows an audit of its results enabling them to support examination and decision making in patent cases. This pilot evaluation has been very promising with results of 87% compared to expert judgement, however, validation in a large scale with practising patent examiners is a major priority for future work.

6.5 Limitations

There are some caveats to be noted. The cross-encoder model used is not a domain-specific model, but rather a generally applicable reranker, that potentially can generate even more improvements in the precision of the relevant rankings for AI healthcare claim structures. The assessment is based on PatentMatch benchmark criteria, which may not be comprehensive enough to fully account for the intricacies of AI in healthcare patents. The outputs of the RAG module are based on retrieved evidence but can be reliant on the LLM, and may not always be consistent. Finally, the corpus is restricted to USPTO filings and it would be greatly expanded to broaden the evaluation scope and generalisability of the findings if other major jurisdiction filings were included as well.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

This work fills the gap between lexical transparency (lexical-based) and the opaque neural semantics that often underpins medical AI tasks, by operationalizing an end-to-end pipeline for Retrieval-Augmented Generation (RAG) in medical context: health-care patent analysis. The deep architecture combines in a meaningful way a sparse vector, representing weak semantics, and a dense vector, representing strong semantics, over an 18,109 document corpus to better utilize complementary linguistic signals for meaningful improvement of baseline recall. However, a second cross-encoder phase greatly improves the top- k accuracy to 0.86 for nDCG@10 and 0.79 for MAP. The final RAG module is based on these high-quality retrievals and generates structured, evidence-based similarity assessments, similar to the verdicts of experts in patent applications, for just about 87% of the cases studied in the pilot. Finally, this holistic solution provides a scalable and easily interpretable solution for prior-art search, without requiring any domain-specific fine-tuning or even domain-specific search spaces.

7.2 Future Work

Future studies should focus on domain-adaptive fine-tuning of the Cross-encoder to further leverage highly complex patent structure that can be captured in such sentences, while investigating learnable fusion mechanisms for jointly optimizing the combination of sparse and dense signals. International filings such as EP, CNIPA, etc. will enable more powerful cross-jurisdictional tracking capabilities. To ensure practical applicability of the RAG outputs, these must be extensively validated by practitioners before they are used as an interactive industry tool. Lastly, the current pipeline does not need fine tuning for an application domain, which makes it very promising for immediate transfer to other data-poor settings such as scientific literature reviews and legal cases.

BIBLIOGRAPHY

- [1] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.
- [2] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, “A guide to deep learning in health-care,” *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.
- [3] L. Jiang *et al.*, “Health-llm: Large language models for health,” 2023.
- [4] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, “Publicly available clinical bert embeddings,” in *Proceedings of the Clinical Natural Language Processing Workshop*, 2019.
- [5] M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, *Current Challenges in Patent Information Retrieval*. Springer, 2013.
- [6] K. Younge and J. Kuhn, “Patent-to-patent similarity: A vector space model,” *SSRN Electronic Journal*, 2016.
- [7] S. Arts, B. Cassiman, and J. C. Gomez, “Text matching to measure patent similarity,” *Strategic Management Journal*, vol. 39, no. 1, pp. 62–84, 2018.
- [8] L. Helmers, F. Horn, F. Biegler, T. Oppermann, and K.-R. Müller, “Automating the search for a patent’s prior art with a full text similarity search,” *PLOS ONE*, vol. 14, no. 3, p. e0212103, 2019.
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT 2019*, 2019.
- [10] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense passage retrieval for open-domain question answering,” in *Proceedings of EMNLP 2020*, 2020.
- [11] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of EMNLP-IJCNLP 2019*, 2019.

- [12] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with gpus,” 2017.
- [13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” 2020.
- [14] M. Douze, H. Jégou, J. Johnson *et al.*, “The faiss library,” 2024.
- [15] R. Nogueira and K. Cho, “Passage re-ranking with bert,” 2019.
- [16] J. Song, C. w. Jin, W. Zhao, A. McCallum, and J.-Y. Lee, “Comparing neighbors together makes it easy: Jointly comparing multiple candidates for efficient and effective retrieval,” in *Proceedings of EMNLP 2024*, 2024.
- [17] M. Suzgun, L. Melas-Kyriazi, S. K. Sarkar, S. D. Kominers, and S. M. Shieber, “The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications,” 2022.
- [18] J. Risch, N. Alder, C. Hewel, and R. Krestel, “Patentmatch: A dataset for matching patent claims & prior art,” 2020.
- [19] R. Whalen, A. Lungeanu, L. DeChurch, and N. Contractor, “Patent similarity data and innovation metrics,” *Journal of Empirical Legal Studies*, vol. 17, no. 3, pp. 615–639, 2020.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30.
- [21] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” 2023.
- [22] M. D. Smucker, J. Allan, and B. Carterette, “A comparison of statistical significance tests for information retrieval evaluation,” in *Proceedings of CIKM 2007*, 2007.
- [23] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques,” *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.
- [24] H. Bekamiri, D. S. Hain, and R. Jurowetzki, “Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert,” *Technological Forecasting and Social Change*, vol. 206, p. 123536, 2024.

- [25] —, “Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert,” 2021.
- [26] R. Krestel, R. Chikkamath, C. Hewel, and J. Risch, “A survey on deep learning for patent analysis,” *World Patent Information*, vol. 65, p. 102035, 2021.