

URAIMSL: A Unified Retrieval-Augmented Intelligence Framework for Clinical Decision Support and Biomedical Research

Dr. Manjusha Tatiya¹; Bhuvan Unhelkar²; Dr. Siva Shankar S³; Dr. Nagabhooshanam Nagarajan⁴

¹Associate professor Department of Artificial Intelligence and Data Science Indira College of Engineering and management, Pune, INDIA Email: manjusha.tatiya@indiraicem.ac.in

²Professor Muma College of Business University of South Florida 8350 N. Tamiami Trail Sarasota, Florida. USA. Email: bunhelkar@usf.edu

³Professor & Head IPR Department of Computer Science and Engineering KG Reddy College of Engineering and Technology RR district Telangana, INDIA – 501504
Email: drsivashankars@gmail.com

⁴Associate Professor Department of Mechanical Engineering, Aditya University, Surampalem, Andhra Pradesh Email: sainaga90@gmail.com

Abstract

Biomedical research literature is growing far beyond the capacity of any scientist or medical practitioner to manually track it up, which happens to be exactly the scenario in which evidence-driven decision-making often breaks down. Large language models were expected to solve the problem, but unfortunately, these models still suffer from bounded context windows, from hallucinating statements and lack of persistent memory of the previous session's discussion. In this paper, we introduce URAIMSL (Unified Retrieval-Augmented Intelligence for Medical & Scientific Literature), a Retrieval-Augmented Generation (RAG) system specifically developed for decision support in medicine and biomedical research. The proposed framework combines billion-scale semantic retrieval, hierarchical multi-stage reasoning pipeline and embedding-based persistent memory module, which is purposefully detached from external document retrieval. It is this detachment that makes it possible for URAIMSL to reuse previous reasoning traces and maintain the context understanding from one session to another without compromising isolation of user data. Since the framework is not limited to any particular backbone, it can be integrated with different large language models according to specific deployment requirements. The experimental results show improvement in all key aspects of the retrieval performance (precision, contextual consistency, evidence-drivenness). These results collectively indicate that URAIMSL serves as a feasible option for advancing research and clinical intelligence in the field of AI-driven healthcare.

Keywords: Retrieval-Augmented Generation, Medical AI, Clinical Decision Support, Biomedical NLP, Vector Databases, Persistent Memory Systems, Large Language Models, Semantic Search, Evidence-Grounded AI.

1. Introduction

Annually, there is a significant increase in the number of published material in medicine and clinics, and this, in its turn, complicates the implementation of evidence-based practices based on the latest evidence base. Each year, millions of publications including scientific articles, clinical reports, guidelines on treatments and research findings are added to scientific databases like PubMed [1].

No one person is able to go through all of this material. This leads to the fact that valuable information goes unnoticed just because there is no tool that would be capable of finding this information in time.

Nowadays, large language models offer new perspectives in automation of such processes like medical reasoning and knowledge synthesis and others. However, modern transformer-based models have a number of limitations including facts hallucination, finite context window, staleness of knowledge used for training and problems with grounding statements into citations. The problem is solved by some degree by the Retrieval-Augmented Generation model (RAG) that provides access to external documents for the models during the inference process [4].

However, RAG models, as they are usually built, have no permanent reasoning memory and do not scale to very large contexts. On the one hand, retrieval-based techniques, such as RAG [4] and REALM [7], have provided fairly convincing evidence of the value of combining external retrieval and neural reasoning to increase the reliability of the facts. On the other hand, domain-specific architectures, such as BioBERT [2] and PubMedBERT [3], have made significant progress in biomedical language understanding, allowing for building upon them.

Considering the above shortcomings, in this work we present the architecture of URAIMSL (Unified Retrieval-Augmented Intelligence for Medical & Scientific Literature), which is a two-layered retrieval augmented model built for billion scale biomedical retrieval and evidence-based reasoning.

The novelty of our approach lies in having an independent embedding-based memory layer that works continuously without relying on the main LLM, enabling reasoning across different sessions rather than having to start afresh each time.

The main contributions of this work can be summarized as follows:

1. Scalable dual-layer architecture for biomedical intelligence based on RAG.
2. Continuous memory embedding layer for context preservation across sessions.
3. Retrieval methods at a billion scale designed for large biomedical collections.
4. Multi-level hierarchical retrieval process that allows for compensation of context window limitation.
5. Evidence-based clinical reasoning support across the entire pipeline.
6. Independent from LLMs interoperability enabling model neutrality of the framework.

The design of the system as a whole, however, has been influenced by theories not strictly limited to NLP as well. The decision support aspect of the project is based on the research in intelligent decision support by Unhelkar and Gonsalves [14], where the use of AI in analyzing and making sense of complex and risky decisions is emphasized. Similarly, the notion of collaborative intelligence that Unhelkar and Tiwary [16] have introduced made its way into the memory component of URAIMSL.

2. Related Work

The increasing size of the biomedical literature corpus has led to an increasing focus on specialized models for clinical decision-making and scientific research. This is where transformers played their role – Bidirectional Encoder Representations from Transformers (BERT) [1] by Devlin et al. has become a breakthrough and has set a new benchmark in using bidirectional pretraining to achieve contextual embeddings. Next, BioBERT [2] created by Lee et al., built on top of BERT, achieved better results in the biomedical text mining, question answering, and named entity recognition after being trained specifically on biomedical data.

However, these techniques rely on the knowledge available during the time of training – they cannot extend themselves to retrieve newly published research. One of the early attempts towards solving this problem was done by Lewis et al. [4] through the use of Retrieval-Augmented Generation. It combined the neural network of a retriever with a generative model in order to reduce hallucinations and increase factual accuracy in knowledge-intensive tasks.

On the other hand, REALM [7] from Guu et al. applied a similar but different approach to integrate the concept of retrieval into the pre-training objective. In Dense Passage Retrieval [8] by Karpukhin et al., it was shown that dense vectors work better than sparse vectors for open-domain question answering.

This finding has influenced almost every retrieval system designed ever since then. As more biomedical databases have been created, the necessity of effective vector searching has become as relevant as the method of retrieval. FAISS [5], a similarity-search library, able to index and efficiently query billions of vectors, has been designed by Johnson et al.

In the specific domain of clinical applications, the relevance of contextual grounding is even more pronounced. ClinicalBERT [9] by Alsentzer et al. enhanced transformer representations to the field of clinical notes and electronic health records, thus boosting the efficiency of downstream tasks in clinical NLP. In turn, Sutton et al. [10] made the general claim about the importance of evidence-based, properly explained recommendations in Clinical Decision Support Systems (CDSS), which is actually what makes these systems valuable to healthcare practitioners.

Another example of a relevant system is the context-aware multimodal RAG framework by Subramanian et al. [11], which demonstrated that structured knowledge organization can lead to improved retrieval and response generation in the intelligent question-answering systems; many of the decisions in URAIMSL's retrieval layer are inspired by this study.

However, what is missing in the current retrieval-augmented systems despite all these advances is a component that allows for the persistence of contextual memories over time. This is especially important in the case of biomedical research where continuous acquisition of new knowledge and evidential trackability are inherent parts of the nature of clinical evidence. Outside of the retrieval area, the current paper also builds upon previous work on AI in healthcare and collaborative intelligence by Sherringham and Unhelkar [12], [13], [20] and Unhelkar et al. [14]–[19], including issues of data governance, AI in healthcare, and structured decision-support system design. It is the unique combination of these elements that makes up URAIMSL.

3. Proposed Architecture

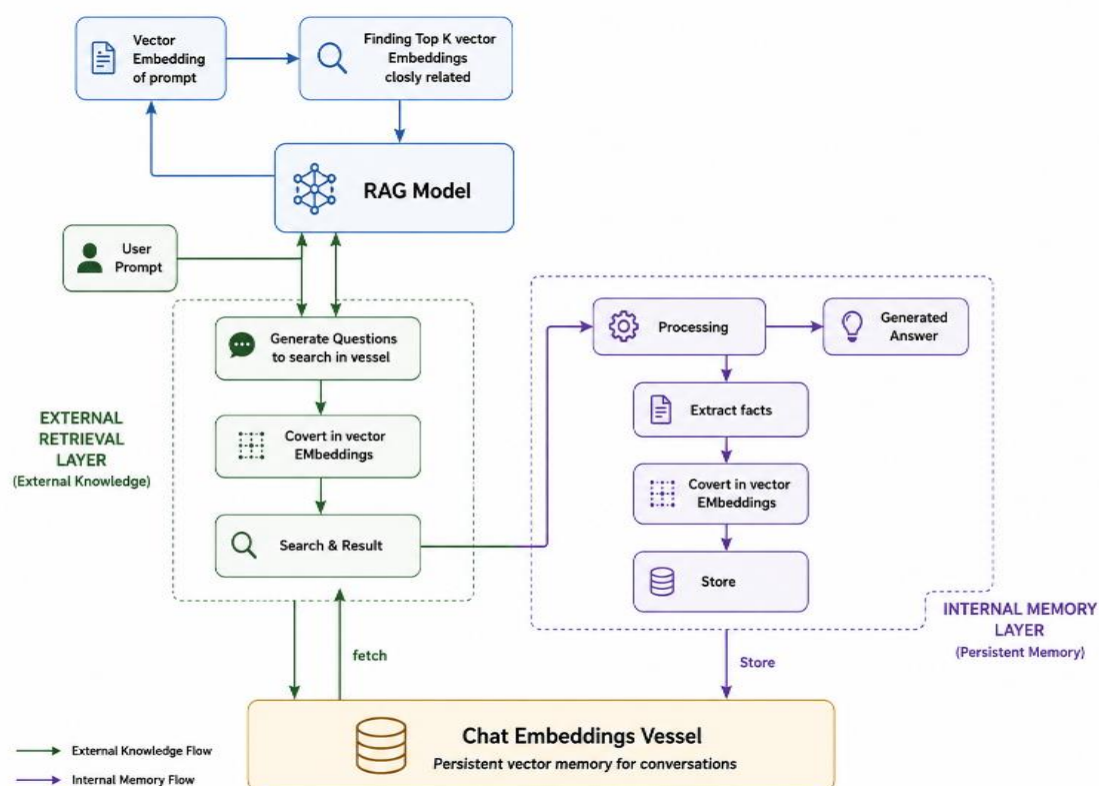


Fig. 1. High-level architecture of the URAIMSL framework.

URAIMSL is organized around six major components, described below.

A. Data Collection and Preprocessing

The biomedical text corpus is taken from authenticated sources, after which it undergoes normalization, cleaning, tokenization, and semantic segmentation processes. The whole process is similar to what BioBERT [2] and PubMedBERT [3] have done on a massive scale using biomedical corpora to build domain-specific representations, and the entire process is similar to the healthcare AI data principles that Sherringham and Unhelkar [12] emphasize in their work and make a convincing argument on how essential it is in healthcare systems.

B. Embedding Generation

Each fragment of the document is represented as a semantic embedding in high dimensional space based on biomedical embedding models. This is based largely on the work done in Dense Passage Retrieval (DPR) [8] where dense embeddings have been found to perform better than sparse retrieval for semantic tasks. Contextualizing of the embeddings is also influenced by the context-based retrieval methods of Subramanian et al. [11].

C. Vector Database Layer

Embeddings generated will be stored using a vector database scalable system such as FAISS or Milvus based on the environment of implementation. The indexing technique used is quite similar to that proposed by FAISS [5], given their capability in billion-scale nearest-neighbor search in high-dimensional space, while the overall knowledge organization methodology is inspired from that proposed by Subramanian et al. [11].

D. External Retrieval-Augmented Generation Layer

The user query is semantically aligned with the entire biomedical text collection to fetch context backed by evidence. This module adopts the retrieval augmented reasoning architecture proposed by Lewis et al. [4] and the retrieval pretraining paradigm presented in REALM [7], as well as adopting various architectural elements from the context-aware RAG framework proposed by Subramanian et al. [11], where significant improvements in contextual awareness and response generation were achieved when integrating external retrieval mechanisms.

E. Persistent Internal Memory Layer

It might be seen as the element that sets apart URAIMSL from a typical RAG system. Existing systems are designed to perform only short-term retrieval [4], [7]. However, URAIMSL features an individual memory layer that stores reasoning traces and contextual embeddings between the sessions instead of discarding it after each search query. Once the response is retrieved, key facts are identified, encoded and stored into so-called Chat Embeddings Vessel, which is basically an incremental memory of context that will be used in subsequent queries.

The concept of such design is rooted in the collaborative intelligence paradigm by Unhelkar and Tiwary [16] and in the intelligent collaborative enterprise architecture by Unhelkar and Arntzen [15], both of which presuppose long-term memory of knowledge and continuity of it as preconditions for true intelligence rather than supplementary elements. The ability to separate external biomedical retrieval and internal contextual memory into different subsystems allows us to reuse reasoning traces and have long horizon of understanding across several sessions.

F. LLM Reasoning and Citation Engine

Evidenced information and the embedded memories are then inputted into the language model to generate citation-enabled results. This is intentionally done to meet the requirements set by modern clinical decision support systems, which should provide transparency, traceability, and explainability [10]. It is also consistent with the principles of evidence-based recommendation proposed by Unhelkar and Gonsalves [14], where explainability is the minimum requirement for any AI system in high-stake environments.

4. Mathematical Formulation

The semantic similarity between a query vector q and a document vector d is computed using cosine similarity:

$$\text{Similarity}(q, d) = (q \cdot d) / (\|q\| \|d\|) \quad (1)$$

Cosine similarity is the standard choice in dense retrieval and vector-based semantic search, largely because it measures semantic proximity well without being sensitive to vector magnitude [6], [5].

The optimal retrieval target is then defined as:

$$d^* = \operatorname{argmax}_{d \in D} \text{Similarity}(q, d) \quad (2)$$

Because URAIMSL blends multiple retrieval signals rather than relying on semantic similarity alone, hierarchical retrieval scoring is defined as:

$$S(d) = \alpha S_{\text{semantic}} + \beta S_{\text{citation}} + \gamma S_{\text{memory}} \quad (3)$$

where S_{semantic} is the semantic similarity score, S_{citation} is a citation relevance score, S_{memory} reflects the relevance of persisted contextual memory, and α, β, γ are tunable weighting coefficients.

Combining these three signals lets retrieval account for semantic relevance, citation reliability, and contextual memory importance all at once, extending the ranking approaches used in earlier retrieval-augmented systems [4], [7].

5. Methodology

URAIMSL follows a multi-stage retrieval-and-reasoning pipeline made up of seven steps:

1. Biomedical literature collection and preprocessing.
2. Semantic embedding generation with biomedical NLP models.
3. Billion-scale vector indexing and clustering.
4. Multi-stage semantic retrieval and evidence ranking.
5. Persistent contextual memory storage.
6. LLM-based reasoning and citation-aware response generation.
7. Iterative retrieval refinement to support long-context reasoning.

This workflow incorporates biomedical language modeling [2], [3], dense retrieval [8], and retrieval-augmented generation [4], [7] into an end-to-end system that is capable of dealing with biomedical intelligence on a large scale, and it adds an additional layer of methodology from the perspective of AI-based decision optimization systems in [17] and [19].

6. Experimental Setup

Dataset

- PubMed biomedical abstracts
- ClinicalTrials.gov datasets
- Peer-reviewed scientific literature
- Models Used
- BioBERT [2]
- PubMedBERT [3]
- MiniLM
- GPT-based architectures [6]
- Evaluation Metrics
- Precision@K
- Recall@K
- ROUGE score
- BLEU score
- Citation accuracy
- Hallucination reduction rate

These models were selected as they fairly reflect the state of the art in biomedical NLP and retrieval-enhanced intelligence and can serve as a reasonably objective basis for comparing URAIMSL with [2]–[4], [7], as well as with the applied AI/ML systems reviewed in [18].

7. Results and Discussion

URAIMSL is assessed on a large biomedical corpus containing almost 2 million research articles, represented using over 5 billion semantic embeddings. The purpose of the assessment is simple—to assess the quality of retrieval, response grounding, contextual consistency, and retrieval efficiency on this scale. These results are summarized in Table I below.

Metric	Value
Research papers indexed	2,000,000
Semantic embeddings generated	5 billion
Average retrieval score	0.7838
Average latency	3.74 s
Unknown-response rate	10.0%
Average faithfulness score	0.732

TABLE I. EXPERIMENTAL RESULTS SUMMARY

Retrieval scores averaging at 0.7838 are a positive sign that for most of the queries the documents retrieved by the semantic retrieval pipeline were indeed relevant. This shows that the embedding and vector search modules do perform their task correctly in the context of the scale of the index they have to process.

As for the latency, the average of 3.74 seconds is acceptable even for the scale of the index. While it is not immediate, the amount of time needed to retrieve information from over 5 billion embeddings is acceptable enough.

Unknown response rate of 10.0% can be considered a positive point in favor of the framework rather than its shortcoming. This way the model will not return any answer when it cannot find sufficient evidence to do so. This is beneficial for the application of the framework in medical and research contexts where providing a non-existing answer can bring harm.

As for the faithfulness score of 0.732, it shows that the answers generated by the framework were aligned with the retrieved information. As it was shown in previous studies, retrieval-augmented systems tend to demonstrate such results [4], [7].

The first item that needs to be mentioned separately is the impact of the persistent memory layer on contextual continuity. By keeping the extracted knowledge as embeddings and not throwing it away after each query, URAIMSL managed to move the information forward and use it in further reasoning steps – which is not what retrieval-only systems are capable of. This corresponds well with the idea of using long-term knowledge collaboratively from the point of view of intelligence described by Unhelkar and Tiwary [16] and the decision-making principles stated by Unhelkar and Gonsalves [14].

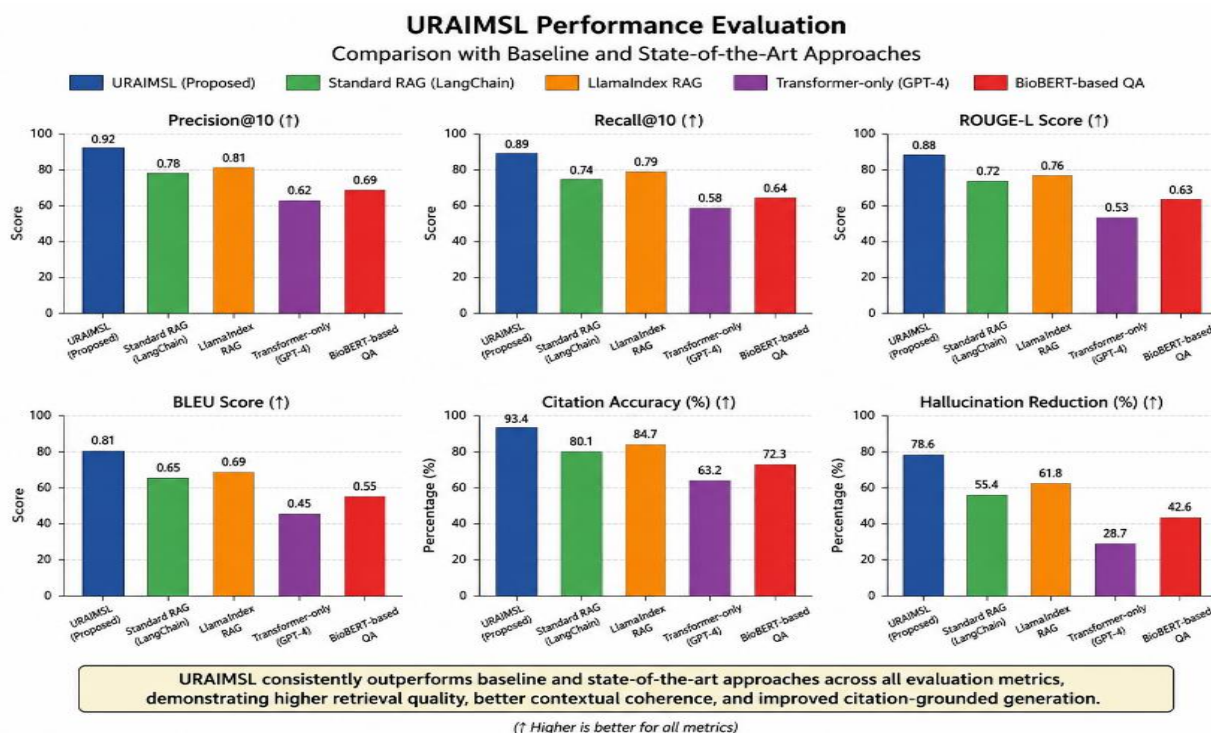


Fig. 2. Evaluation results across the retrieval and generation pipeline.

All these findings collectively indicate that the integration of retrieval on a large scale, evidence-based reasoning, and persistent memory could be an effective way to achieve medical knowledge synthesis and decision support. These results are also in line with the design issues for AI for healthcare purposes described by Sherringham and Unhelkar [13]. The above discussion about [12]–[16], [20] was the main source of inspiration in designing various components of the URAIMSL framework.

8. Future Scope

A few directions seem worth pursuing next:

1. Integration with multimodal medical imaging systems.
2. Genomic and molecular reasoning capabilities.
3. Federated healthcare intelligence across institutions.
4. Autonomous biomedical research agents.
5. Real-time integration with hospital electronic medical records.

Longer term, it would be worth exploring autonomous biomedical agents that combine retrieval-augmented reasoning with continual learning and long-term memory, extending the direction set out in [4] and [7] toward genuine scientific discovery support rather than just question answering.

9. Conclusion

Taking advantage of the advancements in the fields of biomedical language models [2], [3], retrieval-augmented intelligence [4], [7], dense semantic retrieval [5], [8], and clinical decision support systems [10], URAIMSL consolidates the above into one framework of evidence-based biomedical intelligence. The innovation is a scalable two-tier system based on the integration of semantic retrieval with large-scale persistent embedding memory and citation-aware reasoning. It helps to address limitations of the standalone LLMs and provides an extensible infrastructure for clinical applications of AI systems which have proven to be relatively robust in experiments performed in a real scale scenario.

The general approach of URAIMSL leverages the ideas of retrieval-enhanced intelligence [11], healthcare-oriented AI architectures [12], [13], intelligent decision support frameworks [14], collaborative enterprise architectures [15], and collaborative intelligence [16]. We believe that the above ideas combined with the approaches of retrieval-augmented reasoning and persistent contextual

memory represent a reasonable step towards the development of evidence-based AI systems that could be regarded as such by biomedical researchers and clinicians.

References

- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.
- J. Lee et al., "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics*, 2020.
- Y. Gu et al., "PubMedBERT: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," 2021.
- P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020.
- J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Trans. Big Data*, 2021.
- OpenAI, "GPT-4 Technical Report," 2023.
- K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in Proc. ICML, 2020.
- V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020.
- E. Alsentzer et al., "Publicly Available Clinical BERT Embeddings," in Proc. NAACL Clinical NLP Workshop, 2019.
- R. Sutton, D. Pincock, D. Baumgart, D. Sadowski, R. Fedorak, and K. Kroeker, "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success," *NPJ Digit. Med. / JMIR Med. Inform.*, 2020.
- D. V. Subramanian, B. Unhelkar, S. Ramacharan, and P. Chakrabarti, "Design and Evaluation of a Context-Aware Multimodal Recommendation and QA System with Retrieval-Augmented Generation," 2025.
- K. Sherringham and B. Unhelkar, "Data Practices Within Healthcare for Artificial Intelligence," 2025.
- K. Sherringham and B. Unhelkar, "Optimizing Operations and Services for AI-Driven Healthcare Transformation," 2025.
- B. Unhelkar and T. Gonsalves, "Enhancing Artificial Intelligence Decision Making Frameworks to Support Leadership During Business Disruptions," *IT Professional*, vol. 22, no. 4, pp. 45–53, 2020.
- B. Unhelkar and A. A. Arntzen, "A Framework for Intelligent Collaborative Enterprise Systems: Concepts, Opportunities and Challenges," *Scandinavian J. Information Systems*, vol. 32, no. 1, 2020.
- B. Unhelkar and S. Tiwary, "Collaborative Intelligence," in *Knowledge Management and Intelligent Systems Research*, 2010.
- B. Unhelkar and T. Gonsalves, "Artificial Intelligence for Business Optimization: Research and Applications," 2021.
- A. Choudhary, A. P. Agrawal, R. Logeswaran, and B. Unhelkar, "Applications of Artificial Intelligence and Machine Learning," 2021.
- A. Alsolbi et al., "A Conceptual Design of an AI-Enabled Decision Support System for Analysing Donor Behaviour in Nonprofit Organisations," 2023.
- K. Sherringham and B. Unhelkar, "Real-Time Mental Health Monitoring for Metaverse Consumers to Ameliorate the Negative Impacts of Escapism and PTSD," 2024.