**Final Project Report**

**Research Assistant AI: Recommend & Summarize**

## **Abstract**

In the rapidly evolving landscape of scientific research, keeping up with the ever-increasing volume of publications has become a daunting task for researchers. This project presents a hybrid research paper recommendation system that not only suggests relevant academic papers based on an input query but also provides concise summaries to facilitate quicker review. By leveraging Natural Language Processing (NLP) techniques such as sentence embeddings, text summarization, and knowledge graph construction, the system enhances the efficiency of literature discovery. Core technologies used include Sentence Transformers, Retrieval-Augmented Generation (RAG), and Large Language Models (LLMs) like Mistral. The inclusion of graph-based reasoning allows for the identification of thematic clusters, ensuring both semantic and contextual relevance in recommendations. A user-friendly interface was developed using Streamlit to ensure accessibility and interactivity.

## **1. Introduction & Problem Statement**

The exponential rise in research publications poses a significant challenge to researchers aiming to stay current in their fields. Traditional keyword-based search tools often fall short in identifying nuanced relationships between topics, resulting in either information overload or missed relevant works. This project addresses these limitations by developing an intelligent recommendation system that combines the power of embeddings and graph-based reasoning. The goal is to provide users with a set of research papers closely aligned with their interests, derived from a given query or paper title, along with automatic summarization to accelerate the literature review process.

## **2. Literature Review :**

Numerous research efforts have explored content-based and collaborative filtering techniques for academic recommendations. Content-based methods often use TF-IDF, BM25, or newer deep learning-based embeddings to measure similarity. Recent models like SPECTER and SciBERT have demonstrated the effectiveness of sentence-level embeddings for research documents. Graph-based approaches, particularly those involving citation networks and co-authorship graphs, have also proven effective in capturing latent relationships. Retrieval-Augmented Generation (RAG) introduces an innovative hybrid approach by combining retrieval with generative models to enhance contextual understanding. However, most existing systems either rely solely on embeddings or graph structures. This project proposes a hybrid pipeline that integrates both, enriching the recommendation process with structural and semantic depth.

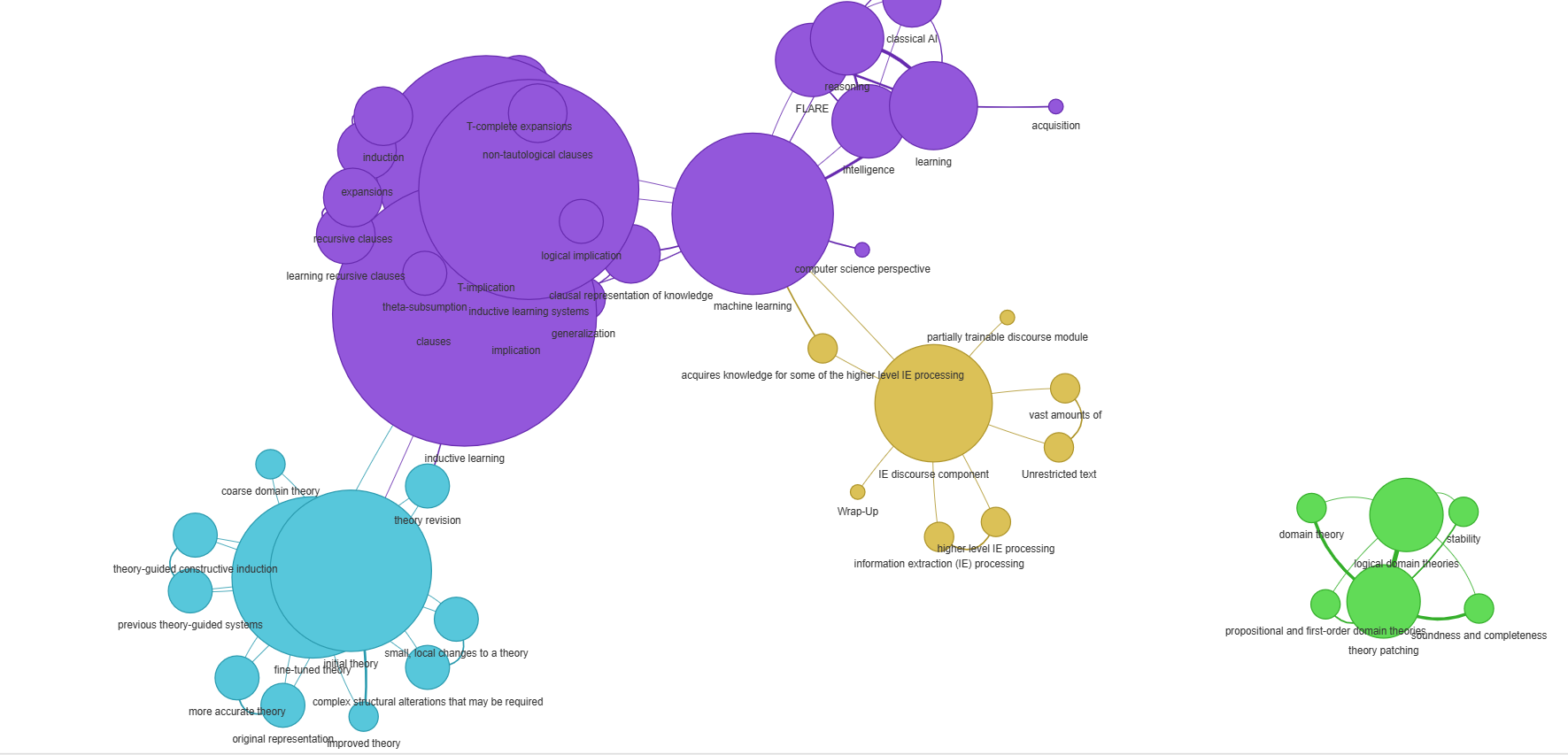
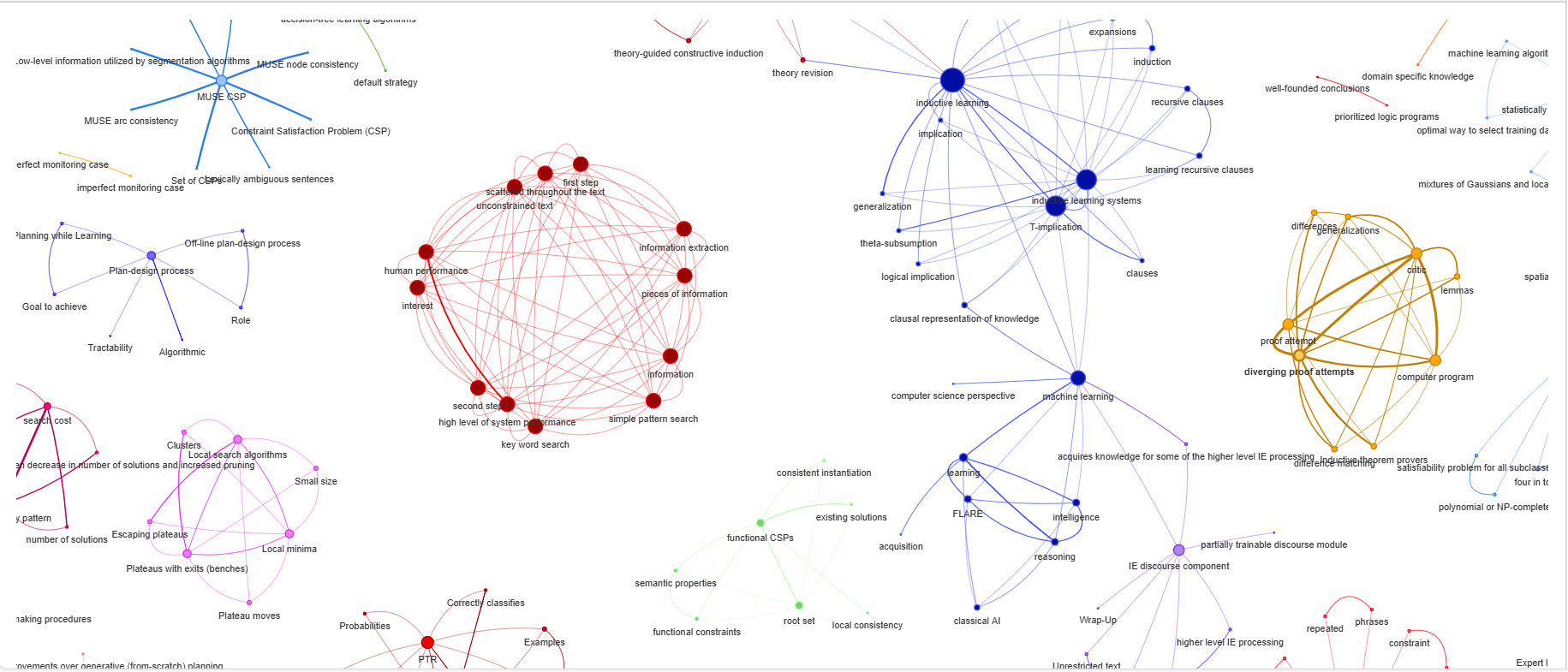
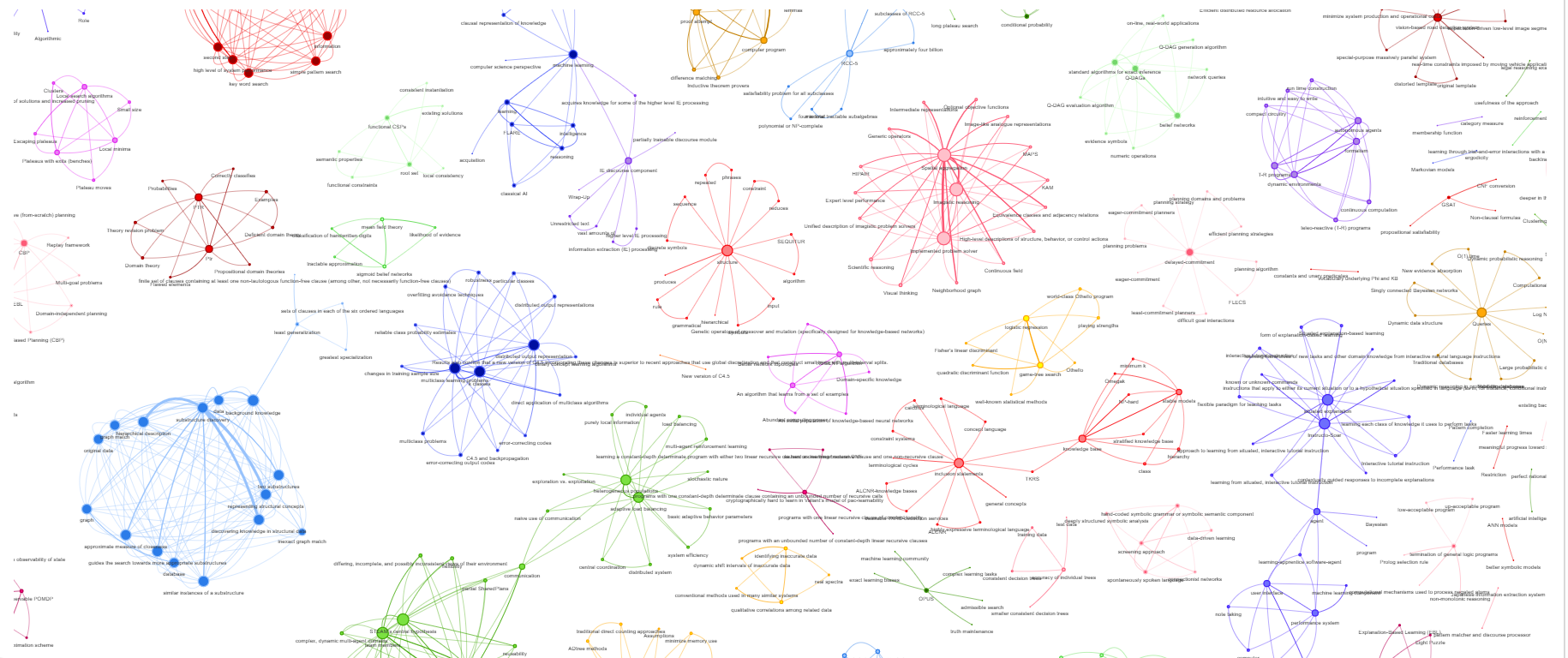
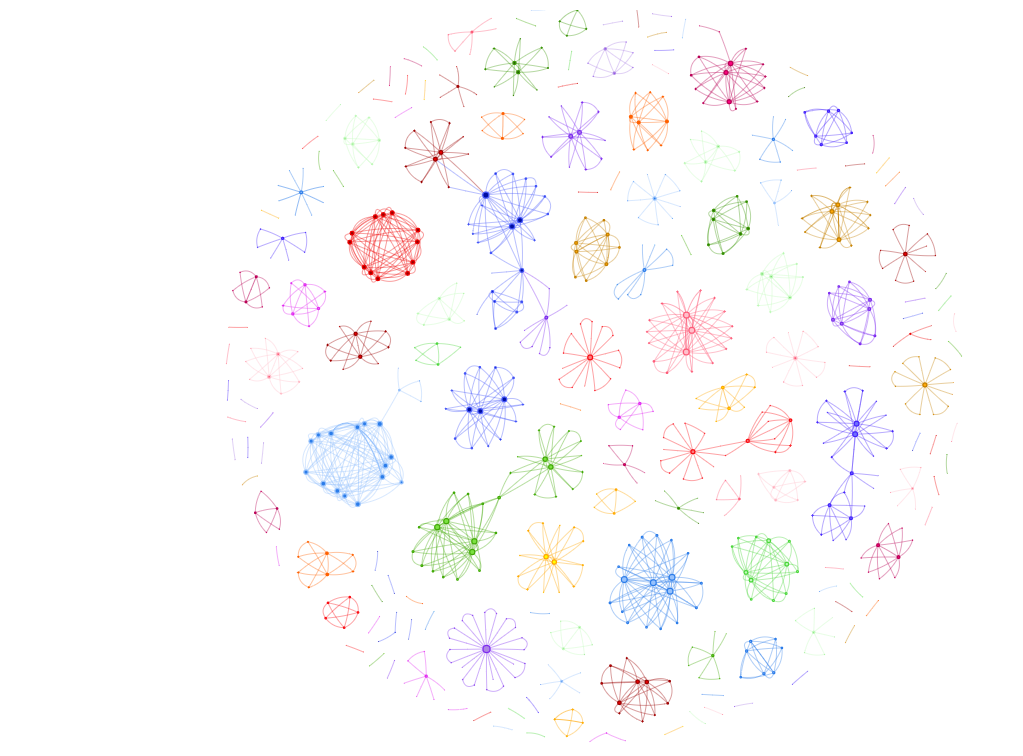
## **3. Methodology**

The proposed system follows a multi-stage approach to recommend and summarize research papers effectively. The process begins with a user providing a query, typically a paper title or abstract. This input is transformed into a dense vector representation using pre-trained Sentence Transformers such as MiniLM or SPECTER2, capturing the semantic meaning of the text. Simultaneously, we utilized the arXiv API to retrieve PDFs of research papers along with their corresponding metadata, including titles, abstracts, and authorship information.

Each abstract from the downloaded papers was passed through the Mistral model via Ollama to extract fine-grained research concepts. These concepts, labeled with categories and importance scores, were computationally expensive to extract due to the depth of analysis required. Using these extracted concepts, a knowledge graph was constructed, where concepts acted as nodes and relationships inferred by the model formed edges. This graph was further enriched using metadata from both the arXiv dataset and a secondary structured dataset (dfg1.csv), which provided topic-based node associations.

Community detection using the Louvain algorithm was then applied to identify clusters of related research themes. This enhanced the recommender system’s ability to find not just semantically similar papers but also thematically related works that might otherwise be overlooked. In the final stage, both the vector-based retrieval and the graph-based communities were leveraged to retrieve a list of top-N recommended papers. Each recommended paper was then passed through an abstractive summarization model (such as T5) to generate a concise, human-readable summary for the end user.

To deliver a seamless user experience, we built a frontend interface using Streamlit. This interface allows users to enter a query and receive real-time recommendations with paper titles, abstracts, links, and summaries.



## **4. Implementation Details**

The system was implemented using Python with support from machine learning and NLP libraries such as PyTorch, TensorFlow, scikit-learn, and HuggingFace Transformers. The initial dataset used was the arXiv metadata CSV (arxivmetadata.csv), which was complemented with a topic relationship file (dfg1.csv). Additional papers and their metadata were dynamically fetched through the arXiv API. This allowed the system to stay up to date and ensured a broader coverage of research domains.

A significant component of the implementation involved processing abstracts using the Mistral model to extract meaningful concepts. This phase proved to be computationally intensive due to the large volume of documents and the depth of semantic analysis involved. The extracted concepts formed the backbone of a custom knowledge graph, which was enriched with real-world citation and topic relationship data. The graph served not just as a visualization of interconnected ideas, but also as a structural component in the recommendation pipeline, enabling community-based filtering and enhanced paper discovery.

Recommendations were generated by combining cosine similarity between query embeddings and paper embeddings with graph-based relevance scoring. This hybrid approach ensured that both direct textual similarity and thematic relationships influenced the final paper selection. For each recommended paper, a text summarization model was used to provide users with an abstract summary, aiding faster comprehension and decision-making.

## **5. Results & Discussion**

One of the notable challenges during implementation was the time-consuming nature of concept extraction and graph formation. Using the Mistral model to semantically analyze and extract concepts from thousands of paper abstracts proved to be the most resource-intensive step. However, the resulting knowledge graph significantly improved the quality and diversity of recommendations.

The integration of community detection within the graph provided additional layers of context, making it possible to recommend papers that were not only similar in terms of content but also relevant within broader research trends. Compared to traditional embedding-only approaches, the hybrid method yielded more interpretable and thematically diverse recommendations. The summarization module further added value by providing concise insights into each suggested paper, enhancing the overall usability and effectiveness of the system.

## **Example Query Recommender output:**

### Finding topics related to: "dynamic backtracking"

Matched Topics: ['Active Learning with Statistical Models', 'Dynamic Backtracking', 'Teleo-Reactive Programs for Agent Control', 'Teleo-Reactive Programs for Agent Control', 'Teleo-Reactive Programs for Agent Control']

Related Topics via Connected Nodes: ['Dynamic Backtracking', 'Teleo-Reactive Programs for Agent Control', 'Active Learning with Statistical Models']

### **knowledge graph recommender:**

Title: Dynamic Backtracking

URL: http://arxiv.org/pdf/cs/9308101v1.pdf

Similarity: 0.4225

Abstract: Because of their occasional need to return to shallow points in a search tree, existing backtracking methods can sometimes erase meaningful progress toward solving a search problem. In this paper, we present a method by which backtrack points can be moved deeper in the search space, thereby avoiding...

--------------------------------------------------------------------------------

Title: Teleo-Reactive Programs for Agent Control

URL: http://arxiv.org/pdf/cs/9401101v1.pdf

Similarity: 0.1138

Abstract: A formalism is presented for computing and organizing actions for autonomous agents in dynamic environments. We introduce the notion of teleo-reactive (T-R) programs whose execution entails the construction of circuitry for the continuous computation of the parameters and conditions on which agent a...

--------------------------------------------------------------------------------

Title: Active Learning with Statistical Models

URL: http://arxiv.org/pdf/cs/9603104v1.pdf

Similarity: 0.0000

Abstract: For many types of machine learning algorithms, one can compute the statistically `optimal' way to select training data. In this paper, we review how optimal data selection techniques have been used with feedforward neural networks. We then show how the same principles may be used to select data for ...

--------------------------------------------------------------------------------

### **Tf-idf recommender:**

Title: Dynamic Backtracking

URL: http://arxiv.org/pdf/cs/9308101v1.pdf

Similarity: 0.513

Abstract: Because of their occasional need to return to shallow points in a search tree, existing backtracking methods can sometimes erase meaningful progress toward solving a search problem. In this paper, we present a method by which backtrack points can be moved deeper in the search space, thereby avoiding...

--------------------------------------------------------------------------------

Title: The Design and Experimental Analysis of Algorithms for Temporal Reasoning

URL: http://arxiv.org/pdf/cs/9601101v1.pdf

Similarity: 0.35

Abstract: Many applications -- from planning and scheduling to problems in molecular biology -- rely heavily on a temporal reasoning component. In this paper, we discuss the design and empirical analysis of algorithms for a temporal reasoning system based on Allen's influential interval-based framework for re...

--------------------------------------------------------------------------------

Title: Efficient Threshold Aggregation of Moving Objects

URL: http://arxiv.org/pdf/cs/0611031v2.pdf

Similarity: 0.149

Abstract: Calculating aggregation operators of moving point objects, using time as a continuous variable, presents unique problems when querying for congestion in a moving and changing (or dynamic) query space. We present a set of congestion query operators, based on a threshold value, that estimate the follo...

--------------------------------------------------------------------------------

Title: Dynamic Large Language Models on Blockchains

URL: http://arxiv.org/pdf/2307.10549v1.pdf

Similarity: 0.145

Abstract: Training and deploying the large language models requires a large mount of computational resource because the language models contain billions of parameters and the text has thousands of tokens. Another problem is that the large language models are static. They are fixed after the training process. ...

--------------------------------------------------------------------------------

Title: Logarithmic-Time Updates and Queries in Probabilistic Networks

URL: http://arxiv.org/pdf/cs/9602102v1.pdf

Similarity: 0.129

Abstract: Traditional databases commonly support efficient query and update procedures that operate in time which is sublinear in the size of the database. Our goal in this paper is to take a first step toward dynamic reasoning in probabilistic databases with comparable efficiency. We propose a dynamic data s...

-------------------------------------------------------------------------------

**6. Conclusion & Future Work**

In this project, we successfully built a hybrid research paper recommendation system that integrates semantic embeddings with a knowledge graph derived from extracted concepts. The combination of LLM-powered concept extraction, graph-based clustering, and sentence embeddings resulted in a robust, context-aware recommender system. Summarization of each recommended paper added another layer of utility, significantly reducing the cognitive load on users.

For future work, improvements can be made in real-time scalability by optimizing concept extraction and graph traversal. Incorporating user feedback loops and reinforcement learning could further personalize recommendations. Additionally, expanding to multi-modal inputs such as figures, tables, and code snippets could elevate the system into a comprehensive research assistant.