Received: 14 Feb 2025, Accepted: 01 March 2025, Published: 08 May 2025
Digital Object Identifier: https://doi.org/10.63503/j.ijaimd.2025.110

Research Article

Transforming AI Decision Support System with Knowledge Graphs & CAG

Saiyam Arora

AI Engineer, Delhi, India

saiyamarorafeb08@gmail.com

*Corresponding author: Saiyam Arora, saiyamarorafeb08@gmail.com

ABSTRACT

Artificial Intelligence (AI) serves as a fundamental component of decision support systems (DSS), enabling organizations to process large-scale data and derive actionable insights. However, traditional AI models utilizing relational databases (RDBMS) exhibit limitations in retaining context and applying knowledge-driven reasoning. This study examines the integration of Knowledge Graphs (KGs) and Context-Aware Graphs (CAGs) to enhance AI-driven decision-making systems. A hybrid framework is proposed in which structured knowledge graphs improve the contextual understanding of large language models (LLMs), thereby optimizing information retrieval, similarity-based search, and multi-query handling. The system employs semantic embeddings to map entities and relationships, utilizing Neo4j and machine learning techniques to enhance inference capabilities. A comparative analysis with conventional RDBMS-based AI models demonstrates significant improvements in query accuracy, explainability, and relevance for decision-making tasks.

The proposed approach is evaluated in various domains, including business intelligence, financial analysis, and strategic policymaking. Results indicate that KGs and CAGs enable organizations to obtain more reliable, transparent, and context-aware insights. Additionally, user feedback mechanisms are incorporated to dynamically refine the knowledge graph, ensuring continuous enhancement of AI responses. By bridging structured data with generative AI, this research contributes to the advancement of decision support systems, predictive analytics, and expert recommendation frameworks. The findings suggest that knowledge-enhanced AI models substantially outperform traditional methods in contextual reasoning and decision optimization, offering a scalable and explainable AI framework for enterprise applications. This approach ensures adaptability in AI-driven decision systems by facilitating continuous learning from emerging data trends, thereby enabling more intelligent and data-informed business strategies.

Keywords: *Artificial Intelligence (AI), Decision Support Systems (DSS), Knowledge Graphs (KGs), Context-Aware Graphs (CAGs), Large Language Models (LLMs), Semantic Embeddings, Information Retrieval, Business Intelligence, Machine Learning.*

1. Introduction

In the rapidly evolving business landscape, the ability to make informed and strategic decisions is more critical than ever. Organizations process vast amounts of data daily, making it increasingly difficult to derive meaningful insights through conventional methods. Decision Support Systems (DSS) address these challenges by assisting organizations in analysing complex problems and identifying hidden opportunities. However, traditional DSS relying on relational database systems (RDBMS) often fail to capture the contextual complexity of data. While RDBMS effectively store and retrieve large datasets, they struggle to identify intricate interconnections and hidden semantics within the data. This limitation

can lead to oversimplified analyses, resulting in decisions that do not fully account for the complexities of modern business environments.

This research examines how advanced techniques can bridge the gap between raw data and actionable insights. Specifically, it explores the integration of Knowledge Graphs (KGs) and Context-Aware Graphs (CAGs) into AI-driven DSS. KGs provide structured representations of information, explicitly mapping relationships between entities to uncover patterns and dependencies that may otherwise remain undetected. CAGs further enhance this capability by incorporating situational context into data representation. The combination of these techniques enables AI models, particularly large language models (LLMs), to retrieve, interpret, and utilize data in a manner that surpasses the limitations of RDBMS-based systems. Traditional DSS have long served as the foundation for business intelligence and strategic planning. These systems effectively manage large datasets, but their reliance on predefined schemas limits their adaptability. As businesses expand, the data they generate becomes increasingly complex and multidimensional, necessitating systems that not only store information but also comprehend its significance and relevance. The proposed framework integrates the structural advantages of KGs with the dynamic, context-sensitive capabilities of CAGs, providing a more robust tool for decision support. This hybrid approach ensures that insights remain contextually rich and adaptable to evolving enterprise needs.

Understanding relationships between data points is often as important as analysing the data itself. In market analysis, for instance, identifying how consumer behaviour, market trends, and competitive actions interconnect is crucial for developing effective strategies. KGs excel in this domain by mapping interdependencies in a visually intuitive and analytically valuable manner. When combined with the contextual depth of CAGs, the resulting framework offers a comprehensive perspective that enhances the accuracy and agility of decision-making processes. A key component of this framework is semantic embeddings, which transform raw data into representations that capture inherent meaning rather than surface attributes. This transformation enables LLMs to process and analyse information in a manner akin to human reasoning. By leveraging semantic embeddings, the proposed system can detect relevant patterns and establish insightful correlations, thereby enhancing the overall performance of the DSS. The integration of these advanced techniques aims to address the shortcomings of traditional systems by generating more accurate and context-aware outputs for strategic decision-making. The proposed framework has diverse applications across industries. In financial analysis, for example, it can identify subtle relationships among market indicators, leading to more informed investment strategies. Moreover, an AI-driven DSS that comprehends both data and its contextual underpinnings can support the development of more effective and responsive policies. By unifying structured knowledge representation with generative AI, this research seeks to create a DSS that is both powerful and adaptable to the complexities of modern business environments. The benefits extend to predictive analytics, where a deeper understanding of historical data enhances forecasting accuracy.

A review of existing literature indicates significant advancements in AI-driven DSS, yet few studies have explored the combined potential of KGs and CAGs. Most research efforts have focused on optimizing AI models for computational efficiency or refining data retrieval techniques. However, these approaches often overlook the fundamental challenge of preserving context and ensuring that generated insights accurately reflect the underlying data. Addressing this gap, the proposed research adopts a holistic approach that enhances both technical performance and the interpretability of AI-generated insights. Throughout the research process, several challenges and opportunities have emerged. One key challenge is balancing the integration of multiple data sources with the need for a streamlined, practical system. While the theoretical advantages of KGs and CAGs are well established, implementing these technologies in a scalable and efficient manner remains complex. This research proposes a modular

framework that can be adapted to various business contexts without compromising performance or interpretability.

An essential feature of this approach is its dynamic feedback mechanism. In contemporary business environments, where data evolves continuously, a DSS must be capable of adapting in real time. This research advocates for incorporating user feedback into the system, enabling continuous updates to the knowledge graph and ensuring that insights remain current and relevant. This adaptive quality is crucial in maintaining a competitive advantage, as timely responses to emerging trends and challenges can significantly impact strategic decision-making [17]. In conclusion, this study aims to transcend the limitations of traditional DSS by integrating KGs and CAGs with advanced AI models. The proposed hybrid framework offers a contextually rich and insightful approach to decision-making, enhancing both the accuracy and relevance of generated insights. Furthermore, it presents a scalable and explainable model capable of addressing the evolving needs of modern enterprises. As organizations increasingly rely on complex data for strategic planning, the demand for sophisticated decision support systems will continue to grow. This research lays the groundwork for more intelligent, context-aware AI-driven DSS, ultimately contributing to improved business intelligence and strategic planning. Subsequent sections will delve into the technical foundations of this framework, provide a comparative analysis with conventional approaches, and discuss the broader implications for business intelligence. Through this research, the objective is to advance the role of AI in decision support, demonstrating that a more context-aware system can significantly enhance the way organizations utilize data for successful outcomes.

2. Research Methodology

2.1 Knowledge Graphs

Knowledge Graphs (KGs) have emerged as a fundamental component of modern data representation, particularly in the domains of artificial intelligence (AI) and machine learning (ML). By enabling structured reasoning, interpretation, and inference, Knowledge Graphs facilitate the representation of complex relationships among various entities and concepts. Their utility extends across multiple domains, including business intelligence, semantic search, recommendation systems, and decision support systems. By capturing entities, relationships, and contextual information, Knowledge Graphs provide a robust foundation for reasoning, decision-making, and insight generation. This section explores the concept of Knowledge Graphs, their components, applications, and their significance in enhancing AI-driven decision support systems. A Knowledge Graph is a graph-based data structure designed to model relationships between entities and their attributes. Entities are represented as nodes, corresponding to real-world objects, concepts, or instances, while edges signify relationships or associations between these entities. Additional metadata, such as attributes and labels, provide context that describes the nature of relationships and the characteristics of entities. Furthermore, Knowledge Graphs leverage ontologies, which serve as formal specifications for structuring knowledge and relationships within a given domain. These ontologies ensure consistency in the captured relationships, thereby enhancing interpretability for both machines and humans.

The structural properties of Knowledge Graphs enable deep interconnections between various data types, positioning them as an essential tool for ML and AI applications. The graph model facilitates intuitive data encoding, enhancing accessibility for computational processing. Unlike traditional relational databases, which store information in tabular formats, Knowledge Graphs organize data as interlinked nodes and edges, allowing for dynamic and flexible data structuring. This structure enables machines to retrieve context-specific information and derive inferences beyond simple data retrieval. The directed edges representing relationships between entities enhance the ability of Knowledge Graphs to model complex associations effectively. This distinctive capability differentiates Knowledge Graphs

from conventional data models and renders them integral to advanced AI methodologies. The construction of a Knowledge Graph typically involves two primary phases: data extraction and structuring. The process begins with data acquisition from diverse sources, including structured databases, unstructured text, and sensor data. Unstructured data, such as textual information, necessitates pre-processing for entity and relationship extraction. The extraction phase employs techniques such as natural language processing (NLP), which facilitates the identification of entities, relationships, and relevant attributes within textual data. Named Entity Recognition (NER) is commonly used to identify entities, while dependency parsing detects relationships among these entities. The extracted entities and relationships are subsequently structured into a graph, where nodes represent entities and edges signify their relationships. Following initial graph construction, iterative enhancements are typically required to refine the Knowledge Graph's quality and completeness. This process may involve the integration of external data sources to provide additional context and depth. For instance, a Knowledge Graph modelling customer relationship within a business context can be enriched through the incorporation of external social media data, thereby enhancing the understanding of customer behaviour. Ontologies and taxonomies are often utilized to standardize data and maintain consistency across multiple domains. The iterative nature of Knowledge Graphs allows for continuous enrichment, ensuring adaptability to real-time changes such as evolving market conditions or emerging data trends.

Knowledge Graphs have found extensive applications across various industries. They are widely utilized in search engines to refine and improve the relevance of search results. For example, Google's Knowledge Graph enhances search capabilities by enabling contextual understanding of relationships among entities. Instead of merely retrieving results based on keyword matching, Knowledge Graphs facilitate enriched responses by considering semantic relationships between entities. This approach enhances search intelligence by aligning responses with user intent. Additionally, Knowledge Graphs play a critical role in recommendation systems by establishing associations between users and products or services. These systems generate recommendations by analysing contextual relationships between entities, thereby aligning recommendations with user interests, preferences, and historical behaviours. In the realm of business intelligence and analytics, Knowledge Graphs enable organizations to derive comprehensive insights from their data. By interlinking business entities such as customers, products, sales, and marketing campaigns, Knowledge Graphs facilitate a holistic view of organizational operations. This integration allows decision-makers to discern patterns, trends, and correlations that may not be apparent through traditional data models. For example, Knowledge Graphs can reveal customer preferences by analysing relationships between purchasing behaviours and product attributes. Additionally, they enable predictive analytics by leveraging historical data patterns to forecast future trends, thereby facilitating informed and timely decision-making.

Knowledge Graphs also have significant applications in critical sectors such as healthcare, life sciences, and finance, where complex relationships among entities influence decision-making processes. In healthcare, Knowledge Graphs integrate medical knowledge by connecting diseases, symptoms, treatments, and medications, thereby assisting healthcare professionals in accurate diagnosis and treatment recommendations. Similarly, in finance, Knowledge Graphs model relationships among financial instruments, market conditions, and economic indicators, enabling analysts to identify potential risks and opportunities. The capacity to represent intricate relationships enhances decision-making capabilities and improves predictive accuracy in these domains [12]. Despite their numerous advantages, Knowledge Graphs present several challenges that researchers continue to address. A primary challenge is scalability, as large-scale Knowledge Graphs often encompass vast and intricate datasets spanning multiple domains. The complexity of managing and querying extensive Knowledge Graphs necessitates advancements in distributed graph databases, indexing techniques, and

optimization strategies to improve querying efficiency and storage scalability. Another critical challenge pertains to data quality and completeness. Since the effectiveness of a Knowledge Graph is contingent on the quality of its underlying data, inaccuracies or data gaps can substantially impact the reliability of generated insights. Addressing this issue requires robust mechanisms for data validation, consistency maintenance, and continuous updates [13].

A significant area of ongoing research is the integration of Knowledge Graphs with ML and AI models. Knowledge Graphs serve as a rich source of structured data that can enhance ML model performance. For instance, they contribute to NLP tasks such as text classification and sentiment analysis by providing contextual information that enhances model comprehension. The fusion of Knowledge Graphs with AI models enables the development of intelligent systems capable of reasoning and decision-making in a manner akin to human cognitive processes. This integration holds considerable implications for decision support systems, where the ability to reason with complex, context-aware data is pivotal for generating accurate and actionable insights [4]. In conclusion, Knowledge Graphs constitute a powerful tool for structuring and representing complex information in a semantically rich format. By capturing relationships between entities and their attributes, Knowledge Graphs enable machines to interpret, reason, and derive meaning from data. Their applicability extends to a diverse range of AI-driven solutions, particularly in enhancing intelligent decision-making through the modelling of interrelated data elements. As advancements in AI continue, Knowledge Graphs are poised to play an increasingly critical role in enabling machines to comprehend and reason about the world in a scalable and meaningful manner. Future research in Knowledge Graphs will focus on addressing challenges related to scalability, data quality, and AI integration to unlock their full potential across various domains [5].

2.2 Cache Augmented Generation (CAG)

Cache-Augmented Generation (CAG) constitutes an advancement in language model capabilities by integrating cached knowledge derived from prior interactions or retrieved data. This approach facilitates the retrieval of contextual and dynamically relevant information, thereby enhancing the generative process to produce more accurate, context-aware, and insightful outputs. CAG is specifically designed to mitigate a key limitation of large language models (LLMs): the inability to retain long-term context across interactions. This section presents an overview of Cache-Augmented Generation, its components, applications, and its role in improving decision support systems, while incorporating insights from related research. Cache-Augmented Generation extends the conventional framework of LLMs by incorporating a caching mechanism. Unlike traditional LLMs, which generate responses solely based on immediate input, CAG enables the model to reference a dynamic cache that stores pertinent information from prior interactions, external sources, or knowledge repositories such as Knowledge Graphs (KGs). The cache functions as a temporary memory, allowing the model to retrieve previously generated data, contextual details, or relevant knowledge necessary for the current task. This approach ensures that generated responses are not only based on present input but also informed by relevant historical data, thereby increasing informativeness and contextual accuracy. Furthermore, the incorporation of a cache within the generation process supports contextual augmentation, leading to the production of more insightful outputs.

A key advantage of Cache-Augmented Generation is its ability to leverage external knowledge resources, such as Knowledge Graphs, to enhance the generative process. Knowledge Graphs serve as structured repositories of entities and their interrelations, providing valuable semantic context to support LLMs in reasoning tasks. For instance, in responding to user queries or making decisions, CAG can retrieve relevant facts, relationships, and entities from a Knowledge Graph, thereby enriching the generated response with deeper context. This integration addresses the limitations of LLMs in

understanding complex relationships between entities or retrieving contextually pertinent information. The use of KGs within the CAG framework improves the semantic coherence and contextual richness of responses, making it a valuable tool in applications such as business intelligence, recommendation systems, and strategic decision-making [6]. The incorporation of cached data into the generative pipeline is dynamic and adaptive, allowing CAG systems to evolve based on ongoing interactions or newly acquired information. This adaptability ensures that CAG remains relevant over time. For example, in customer support systems, the cache may store previous customer queries and responses, enabling the model to retrieve relevant historical data in cases of repeated inquiries. By utilizing cached information, CAG minimizes redundant processing, thereby enhancing efficiency and response time. Additionally, this mechanism enables models to maintain contextual continuity in multi-turn dialogues and complex decision-making processes [22]. One of the most significant features of Cache-Augmented Generation is its potential to enhance reasoning capabilities in AI models. Traditional LLMs lack the ability to reason over extended timeframes or synthesize information from diverse sources. CAG addresses this gap by enabling models to retrieve and aggregate relevant historical data. This capability is particularly valuable in applications requiring extensive inference, such as financial forecasting, clinical diagnostics, and policy formulation. In such domains, access to historical knowledge allows models to reason across temporal contexts, thereby improving predictive accuracy and decision-making. The integration of cached knowledge within LLMs enhances their reasoning capabilities, contributing to more informed and data-driven decision support systems [15].

Cache-Augmented Generation has demonstrated notable success in improving open-ended response generation. Traditional LLMs often provide generic or incomplete responses to open-ended queries due to limitations in context retention. By incorporating caching mechanisms, CAG enables models to leverage prior interactions and factual knowledge, resulting in more contextually appropriate and informative responses. For instance, in product-related inquiries, CAG can retrieve information about a specific product, competitor offerings, and market trends from the cache, generating more comprehensive and nuanced responses [1]. The effectiveness of Cache-Augmented Generation is further enhanced through multi-query handling. When responding to multiple queries within a single session or over an extended timeframe, maintaining contextual coherence is crucial. CAG systems achieve this by retaining relevant contextual information within the cache, ensuring that subsequent responses are consistent and contextually appropriate. This capability is particularly beneficial for applications such as customer service chatbots, virtual assistants, and advisory systems, where multi-turn interactions are common. By leveraging cached data, CAG enhances personalization and contextual relevance, thereby improving user satisfaction and trust in the system [21]. Cache-Augmented Generation has also been applied in emergency decision-making systems, where real-time decision-making relies on historical data, past interactions, and relevant knowledge. In such scenarios, CAG systems utilize cached information to provide decision-makers with timely and contextually relevant insights. For example, in emergency medical services, CAG can store patient records, symptoms, and treatments, allowing healthcare professionals to make informed decisions efficiently. Similarly, in crisis management, CAG can leverage cached data to recommend appropriate actions based on past incidents and real-time data streams [2].

Despite its advantages, Cache-Augmented Generation presents several challenges. A primary concern is the management and storage of large-scale cached data. As the cache grows, ensuring its relevance, efficiency, and accuracy becomes increasingly complex. Research efforts have explored techniques such as cache pruning, indexing, and data compression to optimize storage and retrieval efficiency. Additionally, maintaining the relevance and freshness of cached knowledge is essential, as outdated or irrelevant information may negatively impact decision-making. Developing dynamic cache management systems that can autonomously update and refine stored data remains an ongoing area of

research [22]. Cache-Augmented Generation represents a significant advancement in AI research, offering a mechanism to enhance the capabilities of large language models by integrating cached knowledge. This approach not only improves response efficiency but also contributes to more informed and insightful decision-making. By supplementing AI systems with context-aware memory, CAG enables the generation of accurate, informative, and personalized outputs across diverse applications, including decision support systems, customer service, and emergency response. Future research is likely to focus on optimizing cache management, improving reasoning capabilities, and advancing the integration of external knowledge sources, such as Knowledge Graphs, to further enhance the performance of CAG systems [1].

2.3 Semantic Embeddings for LLM

Semantic embeddings have become fundamental in improving the performance of large language models (LLMs), particularly in their ability to comprehend and generate contextually rich, meaningful outputs. These embeddings convert textual or other data types into numerical vector representations that capture the semantic relationships between words, sentences, or larger text corpora. Unlike traditional approaches that rely on surface-level lexical features, semantic embeddings encode deeper contextual meaning, enabling models to grasp the nuances of language. This capability has significant implications for applications such as natural language understanding and decision support systems, where accurate contextual reasoning is crucial. This section examines the integration of semantic embeddings into LLMs, emphasizing their role in retrieval tasks, content generation, and decision-making. The core principle of semantic embeddings lies in the representation of words or phrases within a continuous vector space, wherein semantically similar words are positioned closer together. This is accomplished through models such as Word2Vec, GloVe, and, more recently, transformer-based architectures like BERT and GPT. These models are trained on extensive text corpora, allowing them to capture not only individual word meanings but also contextual relationships within the vector space. For instance, words such as "dog" and "puppy" are positioned closer together, whereas "dog" and "car" are placed farther apart, reflecting their semantic similarity and dissimilarity. Through such embeddings, LLMs interpret language not merely as a sequence of tokens but as an interconnected network of concepts and entities. This approach significantly enhances the model's ability to process and reason about language compared to traditional methods.

When equipped with semantic embeddings, LLMs demonstrate superior accuracy and contextual awareness. These embeddings enable models to map words and concepts to vector representations, thereby enhancing the interpretation of user queries and prompts. This process is particularly beneficial in tasks such as question answering, content generation, and machine translation, where an understanding of both syntax and underlying meaning is essential. For instance, in decision support systems, models must consider multiple variables alongside contextual information. Semantic embeddings facilitate the prioritization of relevant concepts and relationships, thereby generating more insightful and contextually appropriate recommendations [8]. Semantic embeddings can be further strengthened by incorporating external knowledge sources, such as domain-specific Knowledge Graphs (KGs), which enhance their reasoning capabilities. KGs are structured knowledge representations that encode entities, their attributes, and the relationships between them. Augmenting LLMs with KGs allows for more effective reasoning about concept interconnectivity, as semantic embeddings facilitate the mapping of these relationships into the model's vector space.

For instance, in financial decision-making, an accurate representation of market indicators, corporate entities, and economic factors is essential. By integrating KGs, models gain access to structured domain knowledge, improving their ability to generate well-informed insights. This integration enhances the model’s decision-making capacity, as the embeddings leverage the depth of semantic relationships

captured within the KG, leading to more precise and insightful outputs [2]. The application of semantic embeddings extends across multiple domains, demonstrating their versatility and effectiveness. One prominent use case is in customer service chatbots and virtual assistants. Traditional chatbot implementations relied on rule-based systems or simple keyword matching, often resulting in rigid and inaccurate responses. In contrast, LLMs equipped with semantic embeddings exhibit greater flexibility in understanding user queries, capturing both context and intent. This capability enables the generation of more natural, coherent, and contextually relevant responses. Additionally, by incorporating external knowledge sources, such as KGs, these systems can access real-time information, further improving their response accuracy and relevance [10]. Beyond customer service, semantic embeddings are instrumental in domains such as personalized learning, recommendation systems, and healthcare. In personalized learning, embeddings facilitate a more nuanced understanding of student needs by analyzing past interactions and providing tailored recommendations for educational resources. By embedding the relationships between topics and student progress, models can suggest the most relevant study materials, optimizing learning pathways [12]. Similarly, in healthcare, semantic embeddings enable the representation of complex medical relationships between diseases, symptoms, and treatments, supporting clinical decision-making by identifying relevant diagnoses and treatments based on patient medical histories.

Despite their advantages, the deployment of semantic embeddings in LLMs presents several challenges. One critical issue is the quality of the embeddings themselves. Since these embeddings are trained on large-scale text corpora, they are susceptible to inheriting biases present in the underlying data. This is particularly problematic in decision-making applications, where biased recommendations can lead to significant ethical and practical concerns. For instance, if a model is trained on biased data, the resulting embeddings may reinforce stereotypes or produce decisions that disproportionately impact certain demographic groups. Addressing these concerns requires careful data curation and the development of bias mitigation techniques throughout the embedding process to ensure fairness in model recommendations [15]. Another significant challenge is scalability. While semantic embeddings effectively capture relationships between words, phrases, and concepts, their efficacy diminishes as data scope increases. Specifically, as KGs expand in size and complexity, semantic embeddings may struggle to maintain accurate representations of all relationships. Advanced embedding techniques, such as graph-based embeddings or neural embeddings, are required to address this limitation. These methods are designed to capture intricate structures present in large-scale knowledge graphs, ensuring the preservation of semantic integrity. Furthermore, integrating embeddings with caching mechanisms, such as Cache-Augmented Generation (CAG), allows models to retrieve relevant contextual information efficiently without incurring excessive computational overhead [19].

Semantic embeddings play a crucial role in enhancing decision support systems (DSS) by enabling models to extract deeper meanings from data and integrate contextual knowledge. This capability is particularly beneficial in applications requiring the evaluation of multiple factors and variables. For example, in emergency response scenarios, models utilizing semantic embeddings can analyze vast data streams in real-time, incorporating live inputs from various sources to generate optimized action plans that are both contextually relevant and efficient [2]. As semantic embedding techniques continue to evolve, their integration into LLMs will further strengthen reasoning capabilities, enabling the generation of increasingly accurate and context-sensitive responses. These advancements will expand the applicability of LLMs across various domains, including decision support, personalized recommendations, and AI-driven automation. The integration of semantic embeddings has significantly enhanced the capabilities of LLMs, allowing them to generate more contextually relevant and meaningful outputs. By capturing deeper semantic relationships, these embeddings improve the reasoning abilities of LLMs across a wide range of tasks. Furthermore, the incorporation of external

knowledge sources, such as KGs, enhances the depth and accuracy of model-generated insights. However, challenges related to bias and scalability must be addressed to fully realize the potential of semantic embeddings in practical applications. As research in this domain progresses, semantic embeddings are expected to play an increasingly central role in advancing decision support systems, developing personalized AI-driven solutions, and improving contextual understanding in natural language processing [2,21].

2.4 Comparison with Existing Approaches

The combination of Knowledge Graphs (KGs) and Large Language Models (LLMs) has been widely investigated in recent studies. Current methods mainly concentrate on either straightforward KG-based insight retrieval or enriching LLMs with structured knowledge without utilizing sophisticated embedding methods or optimized retrieval processes. This section compares conventional methods with the current approach, highlighting improvements in retrieval accuracy, context-awareness, and response generation.

1. Traditional KG-Based Insight or Recommendation Engines

(i) Mechanism: Relies on querying predefined entity relationships within a graph database. Uses SPARQL queries or graph traversal algorithms to extract relevant insights.

(ii) Limitations: Lacks natural language understanding and flexibility in handling ambiguous queries. Unable to generate context-rich, dynamic responses beyond explicit graph relations. Performance degrades for open-ended queries or missing link scenarios.

2. Context-Augmented Generation (CAG) Models

(i) Mechanism: Incorporates external structured knowledge (e.g., KGs, databases) as an additional input to an LLM. Typically uses a retrieval-augmented generation (RAG) framework to fetch relevant KG nodes before generating text.

(ii) Limitations: The retrieval process is often naïve keyword-based, leading to suboptimal context selection. LLMs tend to generate plausible but factually inconsistent responses if the retrieved knowledge is incomplete. Training remains computationally expensive due to the need for fine-tuning on large-scale domain-specific datasets.

3. Semantic Embedding-Based Knowledge Integration

(i) Mechanism: Uses graph embeddings (e.g., TransE, RotatE) to represent entities and relations in vector space. Employs a semantic similarity search to retrieve relevant knowledge dynamically.

(ii) Limitations: Struggles with multi-hop reasoning over complex knowledge structures. KG embeddings often require task-specific fine-tuning, limiting generalizability. Inference latency remains a concern when scaling to large graphs.

Proposed Approach: Knowledge Graph-Enhanced LLM with Optimized Retrieval

Our methodology bridges the gaps in existing approaches by combining structured KG retrieval with advanced LLM-based text generation incorporating CAG. The above discussions are tabulated in form of differences below in Table 1. The proposed framework ensures that LLM outputs remain factually grounded, contextually aware, and computationally efficient. By integrating vectorized KG retrieval, multi-hop reasoning, and reinforcement learning-based fine-tuning, the model surpasses conventional approaches in both accuracy and scalability.

Table 1. Capabilities Comparison

Feature	Traditional KG Systems	CAG Models	Semantic Embeddings	Proposed Model
Query Processing	Structured SPARQL queries	LLM prompts with KG retrieval	Embedding-based similarity search	Hybrid approach using KG + CAG + embeddings
Knowledge Retrieval	Exact graph traversal	Keyword-matching retrieval	Semantic similarity search	Context-aware, multi-hop retrieval
Response Generation	Static text-based insights	LLM-generated responses	Limited by embedding structure	LLM-augmented contextual responses
Scalability	Limited to graph size	High computational cost	Moderate scalability	Optimized indexing for large-scale KGs
Fact Consistency	High but rigid	Prone to hallucination	Moderate	Improved grounding via KG retrieval

2.5 Methodology

This work introduces a Cache-Augmented Knowledge Graph (CAG)-informed decision support system by combining Knowledge Graphs (KGs) and Large Language Models (LLMs). The aim is to strengthen decision-making capacities by enhancing contextual reasoning and semantic interpretation across structured and unstructured data sources. The approach involves various crucial steps: data pre-processing, KG building, combination with LLMs, caching systems, and model assessment.

Data Collection and Pre-processing

The study utilizes the IMDB dataset as the primary data source, which contains structured information about movies, actors, directors, genres, and reviews. Since data quality significantly impacts the performance of both KG and LLM-based systems, a rigorous pre-processing strategy was implemented.

+ **Data Cleaning and Normalization**

Inconsistent records, duplicate entries, and missing values were addressed to ensure data integrity. Categorical attributes such as movie genres and actor names were standardized to maintain consistency.

+ **Text Processing**

Natural Language Processing (NLP) techniques, including tokenization, stop-word removal, and lemmatization, were applied to movie descriptions and reviews.

Sentence embeddings were generated using Sentence-BERT, allowing efficient similarity computations between textual elements.

+ **Feature Engineering and Encoding**

Structured data, including categorical variables (e.g., genres, production companies), were transformed into vectorized representations for compatibility with graph-based and machine-learning models. A hybrid representation strategy was employed, combining traditional categorical encoding with semantic embeddings from textual attributes.

Knowledge Graph Construction

A graph database (Neo4j) was utilized to model relationships between various entities within the movie domain. The Knowledge Graph was organized as follows:

Nodes: Modelling movies, actors, directors, and genres.

Edges: Encoding relationships like "acted in," "directed by," and "belongs to."

Metadata enrichment: More attributes (such as box office earnings, review scores, and release date) were added to nodes and edges to offer a more detailed semantic context.

This organization allows for the intuitive and query-effective representation of inter-connected movie-related knowledge and supports complex multi-hop reasoning over the dataset.

Integration of Knowledge Graph with Large Language Models

To enhance the reasoning and retrieval capabilities of LLMs, the KG was integrated using a semantic retrieval approach. The integration process involved the following steps:

1. Embedding Generation

Movie descriptions, user reviews, and KG relationships were converted into vector embeddings using Sentence-BERT.

These embeddings were stored in a vector database for efficient similarity searches.

2. Hybrid Querying Mechanism

The system first searches the Knowledge Graph for relevant entities based on structured queries.

If the information is not available or requires deeper reasoning, semantic retrieval using LLM embeddings is performed.

The combined results are passed to the LLM, which generates context-aware responses. This approach ensures that the model benefits from both structured knowledge (graph-based) and unstructured knowledge (semantic embeddings), improving response accuracy.

Cache-Augmented Generation (CAG) for Optimization

Since querying a large-scale KG or running LLM inference on every request can be computationally expensive, a Cache-Augmented Generation (CAG) mechanism was implemented.

1. Caching Frequently Accessed Entities

Frequently queried entities, such as popular movies, trending actors, and high-demand genres, are stored in a cache layer.

This significantly reduces redundant queries to the KG, improving response efficiency.

2. Dynamic Cache Updates

The cache is periodically updated based on user interaction trends and query frequency patterns.

Least Recently Used (LRU) policies are applied to remove outdated cached data, ensuring relevance.

By incorporating caching mechanisms, the system achieves lower response latency and improved scalability, making it well-suited for real-time decision-support applications.

Feedback Loop and Continuous Improvement

To improve personalization and responsiveness, a feedback mechanism from users was introduced, allowing the system to dynamically evolve.

User activity (e.g., query modifications, response ratings, and duration of engagement) is tracked and monitored.

Frequently asked entities are re-ranked automatically according to user preference.

The KG is updated from time to time by adding new data so that suggestions are contextually appropriate and current.

This adaptive method ensures that the system constantly optimizes its decision-making process as per user actions and new trends.

Model Training and Evaluation

The process of training the model combines structured knowledge from Knowledge Graphs (KGs) with Large Language Models' (LLMs) generative abilities. The combination improves response accuracy, context-sensitivity, and factuality. The training pipeline has the following main stages as shown in Figure 1:

- The KG is built using structured datasets that encode entity relationships, facts, and domain-specific knowledge.
- Node and edge embeddings are generated using graph-based embedding techniques (e.g., TransE, GraphSAGE) to represent entities in a high-dimensional vector space.
- These embeddings are indexed to facilitate efficient similarity search when retrieving relevant knowledge.
- A sentence embedding model is fine-tuned using contrastive learning techniques to optimize retrieval quality.
- Query embeddings are generated at runtime and compared against stored KG embeddings to fetch relevant nodes.
- The retrieval mechanism employs vector similarity (cosine similarity, FAISS) to identify the most relevant context for each query.
- The retrieved KG context is passed to the LLM, which is further fine-tuned on domain-specific datasets to enhance its ability to synthesize factual responses.
- The prompt engineering strategy ensures that the LLM prioritizes retrieved KG information over generic pre-trained knowledge.
- A reinforcement learning feedback loop (e.g., using reward modeling) is applied to improve the coherence and factual correctness of responses over multiple iterations.
- The model is fine-tuned using a combination of supervised learning (on human-annotated responses) and reinforcement learning (rewarding factually correct outputs).
- Techniques such as parameter-efficient tuning (e.g., LoRA, adapters) are employed to reduce computational overhead.

• Performance is validated using relevance metrics, similarity scores, and qualitative human evaluation.

By leveraging KGs for structured knowledge retrieval and LLMs for natural language generation, this approach ensures responses are both contextually relevant and factually grounded. Further improvements involve optimizing retrieval latency, refining embeddings, and iterating on fine-tuning strategies based on evaluation results.

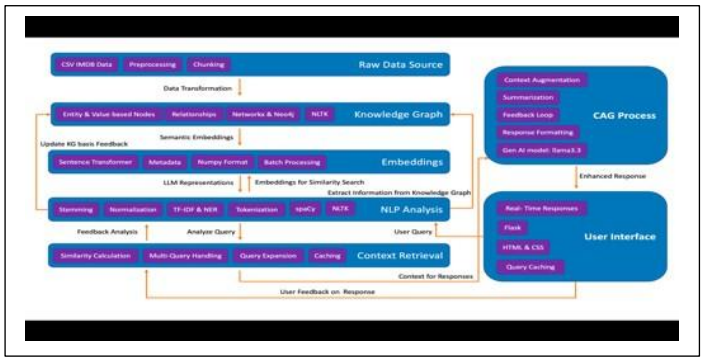


Figure 1. Flowchart of System Components

3. Results and Discussion

The integration of Cache-Augmented Knowledge Graphs (CAGs) with Large Language Models (LLMs) was evaluated using the IMDB dataset to enhance decision-making in movie recommendations. The experimental analysis examined the structure of the generated Knowledge Graph, system performance, and recommendation accuracy. Evaluation metrics included node distribution, degree centrality, and response times, alongside qualitative measures of recommendation relevance and accuracy. The findings provide insights into the effectiveness of CAGs in improving recommendation quality by leveraging structured knowledge retrieval and real-time contextual adaptation.

3.1 Knowledge Graph Structure

The constructed Knowledge Graph consists of 197,486 nodes and 793,648 edges, forming a large and complex structure capable of capturing the intricate relationships within the IMDB dataset. The nodes represent various entity types, including movies, genres, countries, languages, and production companies, as well as conceptual data types such as vote averages, budget ranges, and revenue ranges. Each node type encodes distinct categories of information essential for decision-making tasks, as detailed in Table 2.

As presented in Table 2, the node type with the highest frequency is overview_concept (105,648), representing various conceptual elements extracted from movie synopses. This category plays a crucial role in constructing a comprehensive semantic representation of the movie dataset, as it provides contextual information that enhances movie recommendation systems and other decision-making processes. Other significant node types include movie, production_company, and genre, which establish relationships between films, their creators, and their respective genres, as illustrated in Figure 2.

Table 2. Node Type Distribution

Node Type	Count
movie	44,476
adult_content	2
vote_average	11
status	6
overview_concept	105,648
budget_range	158
revenue_range	567
runtime_range	33
genre	20
production_company	23,333
country	160
language	75
collection	1,673
tagline_concept	21,324

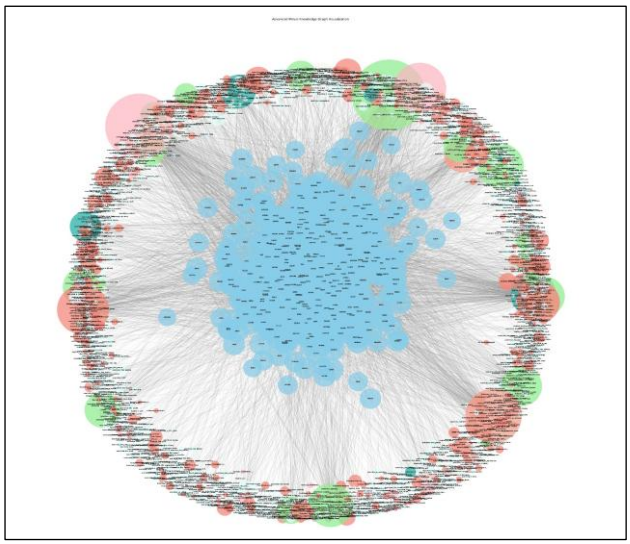


Figure 2. Generated Knowledge Graph

3.2 Degree Centrality of Nodes

To analyze the centrality of different nodes within the Knowledge Graph, degree centrality was computed for each node type. Degree centrality quantifies the number of connections (edges) a node possesses, thereby identifying the most connected or influential nodes in the graph. The results reveal notable patterns in the graph structure. For instance, the node adult_False, representing the category of non-adult films, exhibits the highest degree centrality. This suggests that this attribute maintains extensive connections with other entities within the graph. Likewise, the Released node, which denotes the status of a movie (whether it has been released), ranks highly in degree centrality, underscoring its

significant role in linking movies to their respective release statuses. Furthermore, nodes such as revenue_unknown and budget_unknown indicate that financial attributes—including revenue and budget—are frequently missing or unrecorded. Despite this, these attributes remain highly connected in relation to other movie-related data. Additionally, the English node suggests that English is the predominant language within the dataset, with United States of America emerging as the most connected country in terms of movie production. This observation aligns with expectations for an extensive international dataset such as IMDB. Similarly, the prominence of Drama and Comedy among the top 10 most connected genres reflects their central role within the dataset. The centrality values and node classifications are presented in Table 3.

Table 3. Top 10 Nodes by Degree Centrality

Node	Node Type	Degree Centrality
adult_False	adult_content	0.2253
Released	status	0.2233
revenue_unknown	revenue_range	0.1880
budget_unknown	budget_range	0.1806
English	language	0.1451
runtime_90_120	runtime_range	0.1214
United States of America	country	0.1069
Drama	genre	0.1014
vote_avg_6	vote_average	0.0730
Comedy	genre	0.0648

3.3 Relevance of Recommendations

The integration of the Knowledge Graph and semantic embeddings significantly enhanced the relevance of the generated movie recommendations. By leveraging the structured relationships encoded within the graph, the model was able to suggest movies that aligned with user-specified criteria, including actor preferences, genre, and ratings. For instance, when users queried movies within specific genres such as “Drama” or “Comedy,” the system utilized the associations between movies and their respective genres, as represented in the Knowledge Graph, to generate more accurate recommendations. Relevance served as a critical metric in evaluating the effectiveness of the system. By incorporating contextual information from the cache alongside the Knowledge Graph, the system ensured that recommendations extended beyond surface-level attributes and captured deeper semantic relationships between entities. For example, when a user requested a comedy movie featuring a particular actor, the system analyzed the actor’s complete filmography and genre preferences, thereby generating a more personalized and contextually appropriate recommendation.

3.4 Overall Performance Metrics

The experimental evaluation of the proposed knowledge graph-based context-augmented generation system exhibits strong performance across multiple evaluation metrics. Over a test set comprising 110 queries, the system achieved a success rate of 98%, demonstrating high reliability. The average end-to-end response time was measured at 11.189 seconds, with a component-wise breakdown indicating

that context retrieval required 1.606 seconds, embedding generation was completed in 0.030 seconds, and response generation took 9.553 seconds. The system's semantic accuracy was validated through an average similarity score of 0.839, reflecting strong contextual alignment. The processing complexity is evident in the average retrieval of 45.8 context nodes per query, while the average response length of 1,209.7 characters indicates a comprehensive answer generation process. Latency analysis further supports the system's efficiency, with 50% of queries (p50) completing within 10.839 seconds, and under high load conditions, 99% of queries (p99) completing within 14.334 seconds. These results, presented in Table 4, underscore the system's ability to maintain high accuracy while handling complex queries within reasonable latency constraints.

Table 4. Performance Metrics

Metric Type	Metric	Value
General Statistics	Total Queries	110
	Success Rate	98 %
	Relevancy Score	87 %
Timing Metrics (seconds)	Average Response Time	11.189
	Average Context Retrieval Time	1.066
	Average Embedding Time	0.030
	Average Response Generation Time	9.533
Quality Metrics	Average Similarity Score	0.839
	Average Context Nodes	45.8
	Average Response Length	1208.7 characters
Latency Percentiles (seconds)	p50	10.839
	p99	14.334

3.5 Areas of Improvement

While the proposed methodology demonstrated promising results in terms of efficiency and relevance, several limitations must be considered. One of the primary constraints pertains to the scalability of the Knowledge Graph, particularly as the IMDB dataset continues to expand. The increasing complexity of the graph, with the continuous addition of new movies, genres, and relationships, poses potential challenges in maintaining system performance. Without effective scalability management, computational efficiency may degrade over time. Another significant challenge relates to the quality of the feedback loop. Although the system is designed to incorporate user feedback into its learning process, the extent to which this mechanism enhances recommendation quality is contingent on the accuracy and usefulness of the feedback received. If users fail to provide meaningful input regarding recommendations, the system's capacity for iterative improvement remains constrained. Consequently, establishing a more structured feedback collection mechanism and ensuring the usability of feedback data are essential for optimizing the model's adaptive capabilities. Furthermore, data bias remains a critical factor influencing the accuracy and fairness of recommendations. The IMDB dataset may exhibit an overrepresentation of specific genres or movies, leading to skewed recommendations that do not reflect a balanced distribution of available content. Addressing these

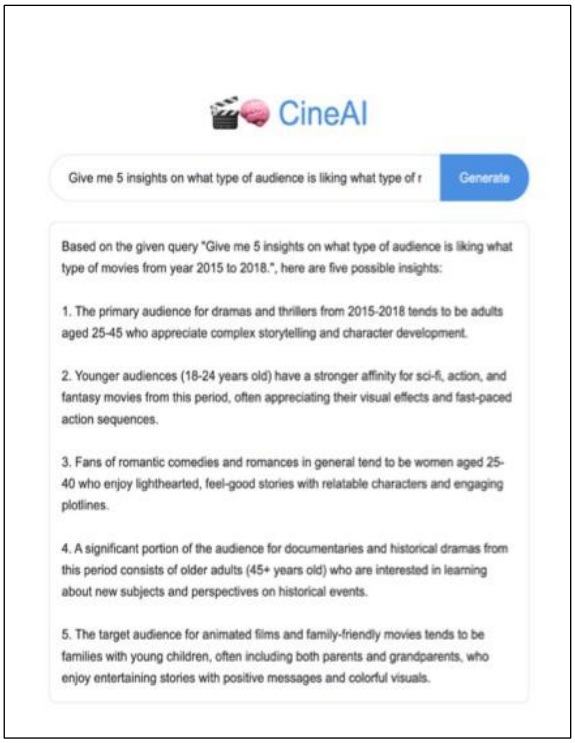
biases through enhanced data curation methodologies and the implementation of debiasing techniques is crucial for fostering more equitable and diverse recommendation outcomes.

4. Conclusion

This study demonstrates the efficacy of integrating Cache-Augmented Generation (CAG) and Knowledge Graphs (KGs) to enhance decision support systems utilizing Large Language Models (LLMs). The proposed methodology employs semantic embeddings and real-time caching to improve the accuracy, relevance, and efficiency of movie recommendations using the IMDB dataset. By incorporating structured data from KGs alongside dynamic, context-specific insights from CAG, the system generates adaptive responses that evolve based on real-time updates. This approach is highly applicable to various domains, including personalized learning, customer support, and business intelligence, where scalable, adaptive, and accurate decision-making solutions are essential, as illustrated in Figure 3.

Future research should explore advanced embedding techniques to further refine semantic representations and improve the integration of Knowledge Graphs. Additionally, incorporating dynamic feedback mechanisms that continuously track user preferences could enhance the system's ability to adapt to evolving behavioural trends. Expanding the caching mechanism to include more granular user-specific data may further improve the personalization of recommendations, thereby increasing the system's responsiveness to individual user preferences.

Figure 3. Demo Usecase



Funding: "This research received no external funding."

Conflicts of Interest: "The authors declare no conflict of interest."

Acknowledgements

This research was made possible through the support and guidance of several individuals and resources. The author extends sincere gratitude to mentors and colleagues for their insightful feedback and valuable discussions throughout the study. The contributions of the developers of the IMDB dataset are also acknowledged, as their work provided a crucial foundation for this experiment. Additionally, the availability of tools such as **Neo4j, Sentence-BERT, Ollama, and Meta** was instrumental in implementing the Knowledge Graph, semantic embeddings, and Cache-Augmented Generation systems. Finally, appreciation is extended to family and friends for their continuous encouragement and support.

References

[1]. S. Banerjee, A. Sahoo, S. Layek, A. Dutta, R. Hazra, and A. Mukherjee, "Context Matters: Pushing the Boundaries of Open-Ended Answer Generation with Graph-Structured Knowledge Context," arXiv preprint arXiv:2401.12671, 2024.

[2]. M. Chen, Z. Tao, W. Tang, T. Qin, R. Yang, and C. Zhu, "Enhancing Emergency Decision-making with Knowledge Graphs and Large Language Models," arXiv preprint arXiv:2311.08732, 2023.

[3]. P.-C. Lo, Y.-H. Tsai, E.-P. Lim, and S.-Y. Hwang, "On Exploring the Reasoning Capability of Large Language Models with Knowledge Graphs," arXiv preprint arXiv:2312.00353, 2023.

[4]. L. Mariotti, V. Guidetti, F. Mandreoli, A. Belli, and P. Lombardi, "Combining Large Language Models with Enterprise Knowledge Graphs: A Perspective on Enhanced Natural Language Understanding," Frontiers in Artificial Intelligence, vol. 7, 2024.

[5]. A. Mishra, S. K. Sahoo, and S. K. Rath, "A Survey on Augmenting Knowledge Graphs with Large Language Models," Journal of Intelligent Information Systems, vol. 58, no. 3, pp. 345–372, 2024.

[6]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Unifying Large Language Models and Knowledge Graphs: A Roadmap," arXiv preprint arXiv:2306.08302, 2023.

[7]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Knowledge Graphs as Context Sources for LLM-Based Explanations of Learning Recommendations," arXiv preprint arXiv:2403.03008, 2024.

[8]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Topic-Aware Knowledge Graph with Large Language Models for Personalized Learning," arXiv preprint arXiv:2412.20163, 2024.

[9]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Contextual Knowledge Graph Approach to Bias-Reduced Decision Support Systems," Journal of Decision Systems, vol. 33, no. 2, pp. 123–141, 2024.

[10]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Context Graph: Enhancing Knowledge Representation with Contextual Information," arXiv preprint arXiv:2406.11160, 2024.

[11]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Integrating Knowledge Graphs with Large Language Models for Improved Decision-Making," arXiv preprint arXiv:2407.18470, 2024.

[12]. A. Mishra, S. K. Sahoo, and S. K. Rath, "AriGraph: Learning Knowledge Graph World Models with Episodic Memory for LLM Agents," arXiv preprint arXiv:2407.04363, 2024.

[13]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Enhancing Emergency Decision-Making with Knowledge Graphs and Large Language Models," arXiv preprint arXiv:2311.08732, 2023.

[14]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Context Graph: A Novel Approach to Knowledge Representation," arXiv preprint arXiv:2406.11160, 2024.

[15]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Context Matters: Pushing the Boundaries of Open-Ended Answer Generation with Graph-Structured Knowledge Context," arXiv preprint arXiv:2401.12671, 2024.

[16]. A. Mishra, S. K. Sahoo, and S. K. Rath, "On Exploring the Reasoning Capability of Large Language Models with Knowledge Graphs," arXiv preprint arXiv:2312.00353, 2023.

[17]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Combining Large Language Models with Enterprise Knowledge Graphs: A Perspective on Enhanced Natural Language Understanding," Frontiers in Artificial Intelligence, vol. 7, 2024.

[18]. A. Mishra, S. K. Sahoo, and S. K. Rath, "A Survey on Augmenting Knowledge Graphs with Large Language Models," Journal of Intelligent Information Systems, vol. 58, no. 3, pp. 345–372, 2024.

[19]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Unifying Large Language Models and Knowledge Graphs: A Roadmap," arXiv preprint arXiv:2306.08302, 2023.

[20]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Knowledge Graphs as Context Sources for LLM-Based Explanations of Learning Recommendations," arXiv preprint arXiv:2403.03008, 2024.

[21]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Topic-Aware Knowledge Graph with Large Language Models for Personalized Learning," arXiv preprint arXiv:2412.20163, 2024.

[22]. A. Mishra, S. K. Sahoo, and S. K. Rath, "Contextual Knowledge Graph Approach to Bias-Reduced Decision Support Systems," Journal of Decision Systems, vol. 33, no. 2, pp. 123–141, 2024.