Journal of Theoretical and Applied Information Technology
28th February 2026. Vol.104. No.4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INNOVATIVE APPROACHES TO EDUCATIONAL KNOWLEDGE GRAPHS: LEVERAGING NEO4J AND LLM FOR SYNTHETIC DATA GENERATION

TANTY OKTAVIA[1], M BAGASKORO TRIWICAKSANA S[2]

[1,2] Information System Management Department, BINUS Graduate Program – Master of Information System Management, Bina Nusantara University, Jakarta Indonesia 11480

E-mail: [1]toktavia@binus.edu, [2]m.triwicaksana@binus.ac.id

ABSTRACT

In today's rapidly evolving higher education landscape, the utilization of data is critical for gaining insights that enhance institutional effectiveness. However, accessing and leveraging this data presents significant challenges due to privacy, licensing, and security constraints. This research evaluates the effectiveness of using Large Language Model (LLM)-generated synthetic data for constructing and visualizing educational knowledge graphs (KGs) using Neo4j. By integrating Neo4j's graph capabilities with LLM's generative power, we developed comprehensive knowledge graphs across five key university domains: curriculum development, personalized learning, learning analytics, resource management, and alumni engagement. Our methodology employs a two-stage prompting strategy with Claude 3.5 Sonnet to generate synthetic datasets that simulate real-world educational scenarios, achieving data completeness rates of 77.55-100% and domain alignment scores of 82.84-100%. The synthetic data successfully captured complex educational relationships through varied node types (5-16) and relationship types (4-16) across domains, while relationship accuracy metrics (38.21-46.55%) demonstrated focused, meaningful connections. Implementation results show efficient processing times (3-8 seconds per domain) and strong practical utility in modelling educational patterns and relationships. This approach significantly reduces data collection time and resources while maintaining data privacy and integrity. While limitations exist regarding validation against real-world datasets, our findings demonstrate that LLM-generated synthetic data can effectively prototype educational knowledge graphs, offering institutions a scalable, privacy-preserving framework for developing data-driven insights and decision-making capabilities.
Keywords: *Knowledge Graphs (KGs), Neo4j, Synthetic Data, Large Language Models (LLMs), Curriculum Development, Personalized Learning, Learning Analytics, Resource Management, Alumni Engagement*

1. INTRODUCTION

The education sector is experiencing an unprecedented transformation, propelled by technological innovations and an increasing demand for efficient, personalized, and data-driven approaches [1]. However, conventional educational management systems face significant challenges in handling the complexity and interconnectedness of modern educational data, particularly in extracting meaningful insights [2]. As the landscape of educational systems continually evolves, numerous challenges emerge that impact students, faculty, and administrative operations alike. From designing coherent curricula and personalizing learning experiences to managing vast amounts of data and fostering alumni engagement, the complexity of these tasks can hinder the effectiveness and efficiency of educational institutions [3]. Knowledge Graphs (KGs) have emerged as powerful tools for representing and analysing complex, interconnected information, making them particularly well-suited to addressing challenges in education. KGs offer flexible, scalable, and context-aware data structures that provide promising solutions to these issues. Unlike traditional relational databases, KGs can easily represent and query intricate relationships between diverse entities within the educational ecosystem [4]. This capability is crucial for developing systems that adapt to the evolving needs of learners, educators, and administrators. Tools such as Neo4j, a leading graph database platform, provide a robust framework for modelling and analysing complex relationships within educational systems. By leveraging Neo4j-based knowledge graphs, educational institutions can seamlessly map and visualize with integrate disparate data

sources, enhance data accessibility, and facilitate intelligent decision-making processes[5]. This integration is pivotal in addressing critical areas such as curriculum development, personalized learning, learning analytics, resource management, and alumni engagement [1].

Educational institutions generate vast amounts of data across multiple domains, including curriculum development, personalized learning, learning analytics, resource management, and alumni engagement. However, the effective utilization of this data presents significant challenges. Privacy concerns regarding student and institutional data, licensing restrictions on educational content, and stringent security requirements for sensitive information create substantial barriers to data accessibility. These obstacles are further exacerbated by data fragmentation across departmental silos, inconsistent data management practices, and challenges in integrating information from disparate sources, ultimately hampering institutions' capacity to leverage data for informed decision-making. Moreover, existing knowledge graph research often lacks transparent documentation regarding data licensing, privacy protocols, and security measures critical considerations given that data quality directly influences the effectiveness and value of knowledge graph implementations [6].

The integration of LLM with graph database technology presents new opportunities to address the data acquisition challenges faced by educational institutions. Synthetic data emerges as a compelling solution to these challenges, offering a way to generate artificial datasets that maintain the essential characteristics and relationships found in real educational environments while eliminating privacy concerns[7]. In the context of education, synthetic data becomes particularly valuable as institutions grapple with strict privacy regulations and limited access to comprehensive datasets that could drive meaningful analytics. The advent of LLM has revolutionized our ability to generate high-quality synthetic data. These models demonstrate remarkable capabilities in understanding and replicating complex patterns within educational contexts, from student behaviours patterns to institutional processes. Their ability to maintain contextual relevance while generating diverse, interconnected data points makes them particularly well-suited for creating educational datasets that can serve as foundations for knowledge graph development. The strength of LLM-generated synthetic data lies in its adaptability and control. Educational institutions can generate specific scenarios, manipulate variables, and test edge cases without the constraints and risks associated with real student data. This flexibility, combined with the consistency and completeness of generated datasets, provides a cost-effective approach to developing and testing educational analytics systems. Furthermore, the ability to generate uniform data structures with controlled noise levels offers unique opportunities for systematic analysis and system development [8].

However, this emerging capability raises important questions about the potential of LLM-generated synthetic data in educational knowledge graph scenarios. A fundamental question arises: How effectively can synthetic data generated by LLM be used as a prototype to develop knowledge graph educational scenarios that accurately reflect real-world relationships and patterns and help gain insights? Effective in the fundamental question means the extent to which the synthetic data generated by LLM can mimic real-world relationships and patterns, provide valid insights, reduce the cost and time to collect real data, and provide opportunities to test various scenarios. This question becomes particularly relevant when considering the complex web of relationships in educational settings, from student, lecture, courses, curriculum, and etc interactions to other educational aspects. By generating synthetic data that preserves the complex relationships and patterns present in real educational scenarios, this approach offers a promising solution to the data accessibility and privacy challenges that often hinder educational analytics initiatives. This ability to generate synthetic data, combined with Neo4j's powerful visualization and graph analysis features to help derive insights from data, sets the stage for convincing educational institutions to develop more comprehensive knowledge graphs in the future that support decision making using real data [9].

This research introduces an innovative methodology for constructing and visualizing educational knowledge graphs using Neo4j, combined with synthetic data generation through LLM using Claude 3.5 Sonnet. The study focuses on five critical domains within educational institutions: curriculum development, personalized learning, learning analytics, resource management, and alumni engagement. The proposed methodology addresses the fundamental challenges of data acquisition and utilization in educational settings while maintaining data privacy and

security. By generating synthetic datasets, educational institutions can simulate real world educational scenarios and relationships without compromising sensitive information. This approach enables the development of rich, interactive knowledge graphs that support advanced analytics and decision-making processes while adhering to privacy regulations and ethical considerations. Through this research, we aim to present a comprehensive review of relevant literature, examining current approaches to educational data management and knowledge graph applications. Subsequently, we detail our methodology for LLM-generated synthetic data and integrating with Neo4j. We then present our implementation results and analyse their implications for educational institutions. Finally, we discuss our findings and outline future research directions in this promising field.

2. THEORITICAL BACKGROUND

2.1 Knowledge Graphs (KGs)

Knowledge Graphs (KGs) have emerged as a pivotal technology in the realm of data representation and semantic understanding. Fundamentally, a Knowledge Graph is a structured representation of information, where entities (such as people, places, concepts) are interlinked through various relationships, forming a network of knowledge. Unlike traditional relational databases, KGs emphasize the connections and context between data points, enabling more intuitive and meaningful data retrieval and inference. The core components of a Knowledge Graph include nodes (representing entities), edges (denoting relationships), and properties (attributes of entities and relationships). This graph-based structure facilitates the integration of heterogeneous data sources, supporting complex queries and enabling advanced analytics. Furthermore, Knowledge Graphs leverage ontologies and schemas to ensure semantic consistency and to provide a framework for data interoperability [5].

The most prominent implementations of Knowledge Graphs are Google's Knowledge Graph, which enhances search engine results by providing structured and relevant information directly within the search interface. Beyond search, KGs are instrumental in various applications, including recommendation systems, natural language processing, and artificial intelligence, where understanding the context and relationships between entities is crucial. In the educational sector, Knowledge Graphs offer transformative potential by enabling personalized learning, enhancing curriculum design, and improving educational analytics. By modelling educational content, learner profiles, and interactions as interconnected entities, KGs facilitate a deeper understanding of the learning process and outcomes. Knowledge Graphs also can enable the creation of adaptive learning systems that tailor educational content to individual learner needs. By analysing the relationships between learning objectives, resources, and student performance data, KGs can recommend personalized learning paths and resources, thereby enhancing student engagement and efficacy [10]. In Curriculum Design and Management, educators and institutions can leverage Knowledge Graphs to design and manage curricula more effectively. By mapping out the relationships between various courses, prerequisites, and learning outcomes, KGs assist in identifying gaps, redundancies, and opportunities for interdisciplinary integration. also in Educational Analytics, KGs provide a robust framework for aggregating and analysing diverse educational data sources. This holistic view enables educators and administrators to gain insights into student performance trends, dropout rates, and the effectiveness of teaching methodologies. Moreover, KGs support predictive analytics, helping institutions proactively address potential challenges in the educational ecosystem [11].

2.2 Neo4j

Neo4j is one of the leading graph database platforms specifically designed to store, manage, and analyse data structured in graph form. As an implementation of the property graph model, Neo4j enables the representation of complex data and interconnections between entities with high efficiency, making it an ideal choice for implementing Knowledge Graphs in various domains, including education. One of the key advantages of Neo4j is its ability to handle complex queries with high performance through the intuitive Cypher query language. This allows developers and researchers to easily extract relevant information from large, interrelated data networks without having to perform complex optimizations as with traditional relational databases [5].

Neo4j enables modelling of relationships between various elements such as learning materials, curriculum, student profiles, and learning interactions. By utilizing Neo4j, educational institutions can develop a more in-

depth analytics system to understand learning patterns, identify individual student needs, and design more effective interventions. Neo4j supports integration with various modern technologies and frameworks, including Machine Learning and Large Language Models. which enables the development of innovative solutions such as synthetic data generation for educational research and development. Neo4j's flexibility and scalability make it capable of handling the rapid growth of data and the increasing complexity of relationships in dynamic educational environments [5]. Neo4j implementation in education is also supported by an extensive ecosystem, including data visualization tools, plugins, and an active user community. This facilitates the adoption and development of graph-based applications that can be customized to the specific needs of educational institutions. Thus, Neo4j not only provides a robust technical infrastructure, but also supports innovation and collaboration in education data management and analysis.

2.3 Synthetic Data

Synthetic data refers to data that is artificially generated using algorithms and mathematical models, rather than collected from direct observation or measurement of real phenomena. Synthetic data is becoming increasingly important in various fields, including education, because it can overcome the limitations of real data such as privacy, limited amount of data, and bias in data. By generating data that resembles real data but without containing sensitive information, synthetic data enables the development and testing of safer and more effective machine learning models [8].

Benefits of Synthetic Data:

- Privacy and Security, synthetic data can be used without revealing personal or sensitive information, reducing the risk of privacy breaches.
- Availability and Scalability, synthetic data can be generated in large quantities as needed, overcoming the limitations of real data that may be difficult to obtain or expensive to collect.
- Reduced Bias, with full control over the data generation process, synthetic data can be designed to reduce or eliminate biases present in real data, improving model fairness and accuracy.

The Leading companies in artificial intelligence development, such as OpenAI, Meta, and Anthropic have adopted the use of synthetic data generated by their Large Language Models (LLMs) to train new models. This approach allows them to improve the model's capabilities while addressing challenges related to real data. OpenAI, under the leadership of Sam Altman, has utilized synthetic data as part of their LLM models training strategy [12]. By using data generated by previous models, OpenAI can expand the training dataset without having to rely entirely on data collected from external sources. This not only speeds up the training process but also enables model testing and validation in various synthetically generated scenarios. Also, Meta has also implemented the use of synthetic data in their LLM development. Meta uses data generated by their LLM to train new models, enabling faster iterations and improved model quality. In addition, the use of synthetic data helps Meta maintain user privacy by reducing the need to use real data that could contain sensitive information. The integration of synthetic data in educational Knowledge Graphs can enrich data representation and improve analytic capabilities. By using synthetic data, Knowledge Graphs can cover a wider and more varied range of learning scenarios, enabling the development of systems that are more adaptive and responsive to individual student needs. In addition, synthetic data can be used to simulate complex interactions between various entities in Knowledge Graphs, improving understanding and prediction of learning dynamics [13].

2.4 Large Language Models (LLMs)

Knowledge Large Language Models (LLMs) are one of the leading innovations in the field of artificial intelligence that focus on natural language processing and generation. Recent advancements in Large Language Models (LLMs) like GPT-4o, GPT-3.5 Sonnet, and o1 represent a significant leap forward in the field of synthetic data generation. These models show a marked improvement in their ability to generate synthetic data that closely mimics real-world conditions, which is valuable in several AI applications such as model training, reinforcement learning, and general AI benchmarking. Large Language Models such as GPT-4o, o1-preview by OpenAI and Claude 3.5 Sonnet by Anthropic have significantly enhanced their capabilities in generating synthetic data that effectively mirrors real-world complexity [14]. GPT-4o has demonstrated its prowess in generating massive, diverse datasets through persona-driven prompts [15]. This approach incorporates contextually relevant personas to guide the data synthesis process, enabling the

model to generate data that not only meets technical requirements but also captures nuanced human-like scenarios. This methodology has proven particularly effective for large-scale data generation, successfully creating billions of diverse personas while minimizing redundancy through advanced deduplication techniques, including MinHash and embedding-based methods. Additionally, Claude 3.5 Sonnet's 200,000-token context window makes it particularly suitable for generating extensive synthetic datasets [14].

The o1 model represents another evolution in synthetic data capabilities by combining reinforcement learning with standard LLM architecture. The reinforcement learning phase in o1 involves training the model to improve the quality of generated data by learning the best reasoning paths through chain-of-thought reasoning, which helps in refining the synthetic outputs. This dual-layer approach not only ensures that the generated data aligns well with expected outcomes but also gives o1 a unique edge in tasks that involve complex, multi-step reasoning processes, crucial for mimicking real-world decision-making scenarios. Impact on Synthetic Data Quality: With these advancements, models like GPT-4o and 3.5 Sonnet are pushing the boundaries of how synthetic data can be used to train and improve AI models. These models can create synthetic datasets that are increasingly indistinguishable from real-world data, especially when persona-driven prompts are used, which helps in emulating specific user scenarios or professional contexts. As a result, LLMs are not only advancing in their core language generation tasks but are also playing an instrumental role in improving data availability for AI training without the need for massive real-world datasets [12].

2.5 Applications of KGs in Educational Domains

2.5.1 Curriculum Development

Knowledge Graphs have shown promise in enhancing curriculum development processes. Researchers have used KGs to analyse the alignment between course content and industry requirements, leading to curriculum updates that increase graduate employability. The graph structure allows for easy identification of skill gaps and redundancies across the curriculum.

2.5.2 Personalized Learning

The application of KGs in personalized learning has been a subject of intense research. Recent studies have developed KG-based recommendation systems that suggest learning resources based on a student's prior knowledge, learning style, and career goals. These systems have shown higher engagement rates compared to traditional recommendation algorithms.

2.5.3 Learning Analytics

In the field of learning analytics, KGs have enabled more sophisticated analysis of student data. Graph-based approaches have been used to identify at-risk students with improved accuracy compared to previous methods. These models typically incorporate not just academic performance, but also social interactions and resource utilization patterns captured in the graph structure.

2.5.4 Resource Management

KGs have also proven valuable in optimizing educational resource management. Studies have demonstrated how the graph structure of KGs allows for efficient mapping and utilization of educational resources, optimizing their allocation and accessibility.

2.5.5 Alumni Engagement

In the domain of alumni engagement, KGs have opened new possibilities for maintaining and leveraging alumni networks. Graph-based approaches have been used to match alumni, Knowledge graphs map alumni networks, track career trajectories, and analyze engagement patterns, providing valuable insights for tailored outreach. This interconnected approach not only strengthens alumni relations but also contributes to the overall growth and sustainability of the institution.

2.6 Related Work

The research paper "Multi-source Education Knowledge Graph Construction and Fusion for College Curricula" addresses the fundamental challenge of managing fragmented learning materials in Electronic Information education at the university level. Students often struggle to navigate through extensive educational content spread across textbooks, slides, and syllabi, which can impede their learning progress. To address this challenge, we propose an automated framework that constructs and fuses multi-source knowledge graphs using Natural Language Processing (NLP) techniques [4].

Traditional approaches to curriculum knowledge organization face two major

limitations: the resource-intensive nature of manual knowledge graph construction and the narrow scope of existing solutions that typically cover only isolated segments of course content. Our framework overcomes these constraints by automating the extraction and integration process, resulting in a comprehensive knowledge graph that spans multiple levels of the Electronic Information curriculum. The key contributions of this study include:

- A novel method for integrating heterogeneous educational resources into a unified, multi-layered knowledge graph.
- Automated identification and visualization of concept relationships across different courses.
- Tools for comprehensive knowledge visualization that benefit both students and educators.

This framework not only helps students grasp the interconnections between different topics but also enables educators to optimize their teaching strategies through better curriculum structure visualization. The research paper "Synthetic data generation methods in healthcare: A review on open-source tools and methods", explore methods for generating synthetic data to address challenges in healthcare, including data scarcity and privacy concerns. The study reviews various tools and techniques for synthetic data creation across domains like tabular, imaging, and omics data. Deep learning-based synthetic data generators are highlighted to lower clinical trial costs, improve AI model accuracy, and protect patient privacy. However, maintaining data fidelity and mitigating biases remain significant challenges. The authors emphasize the need to balance realism and privacy to ensure effective generalization across diverse patient populations [8].

The research paper "Analysis of Learning Effectiveness Based on Knowledge Graph", analyse learning outcomes by constructing a knowledge graph for a "Data Structure and Programming" course, integrating data from platforms such as PTA and MOOCs. The study aims to identify learning difficulties and provide targeted support, enabling more personalized education. While the approach demonstrates potential, the authors acknowledge that current methods lack robust visualization tools and are mostly focused on elementary education rather than the more diverse and complex domains of higher education. The proposed method uses big data and AI to enhance learning experiences and adapt curricula based on student needs[6].

The research paper "Automatic Construction of Subject Knowledge Graph based on Educational Big Data", introduce a bootstrapping strategy for automatically constructing subject-specific knowledge graphs using educational big data. This iterative approach extracts core knowledge points from resources like textbooks and expands the graph by integrating data from internet encyclopaedias. The study employs BERT-BiLSTM-CRF for automatic entity recognition and ensures graph enrichment by evaluating the semantic relationships between nodes. Despite the reduced dependence on manual input, maintaining an accurate and up-to-date representation of subject knowledge remains a challenge, particularly as educational content evolves [11].

The research paper "EDUKG: a Heterogeneous Sustainable K-12 Educational Knowledge Graph", a comprehensive K-12 educational knowledge graph designed to overcome the limitations of previous educational knowledge graphs, which often lacked interdisciplinary scope or failed to evolve over time. EDUKG uses an interdisciplinary ontology with 635 classes and 445 properties to represent educational knowledge and resources uniformly. The knowledge graph is also designed for sustainable maintenance, integrating new information incrementally to adapt to educational changes. EDUKG stands out for its ability to provide rich, adaptive support for educational technologies by dynamically linking a wide range of knowledge and resources [10].

The research paper "Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cybersecurity Education", focus on creating knowledge graphs from unstructured texts to enhance cybersecurity education. Given the absence of standard datasets or pre-defined ontologies, the authors adopt a bottom-up approach that combines machine learning with expert input to establish entity relationships. This method integrates diverse resources, such as lecture notes and online content, to develop a dynamic, interactive learning environment. A key contribution of this study is the construction of a knowledge graph that facilitates hands-on cybersecurity training through visualization, helping students grasp complex attack-defence scenarios more effectively [11].

While existing studies demonstrate the potential of Knowledge Graphs (KGs) in various

domains, they often overlook critical aspects of data management, particularly concerning licensing, privacy, and security issues in data collection and utilization. Our research addresses these limitations by introducing a framework for synthetic data generation using Large Language Models model Claude 3.5 Sonnet and providing a comprehensive evaluation of KGs in educational contexts. This study specifically addresses the following research gaps:

1. Lack of systematic frameworks for generating privacy-compliant synthetic educational data using Claude 3.5 Sonnet, which can simulate diverse learning scenarios while maintaining data quality and relevance.
2. Limited integration capabilities in existing KG frameworks, which typically focus on isolated educational domains rather than providing a unified, cross-domain knowledge representation.
3. Absence of comprehensive evaluation metrics that assess KG performance specifically within educational contexts, considering both technical efficiency and pedagogical effectiveness.

3. METHODOLOGY

This research employs a systematic methodology for constructing educational knowledge graphs using synthetic data generation and graph database technologies. The methodology consists of two primary phases. Synthetic Data Generation and Knowledge Graph Construction & Visualization, followed by a comprehensive evaluation.

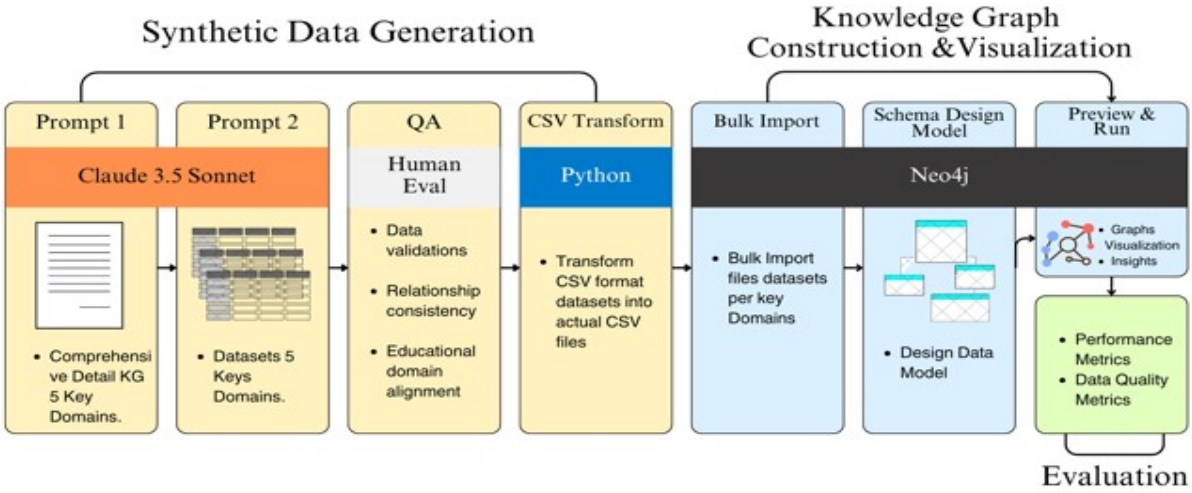

Figure 1: Framework Methodology

3.1 Synthetic Data Generation

During the synthetic data generation phase, we utilized Claude 3.5 Sonnet, a state-of-the-art large language model, to produce synthetic educational data across five key domains. This implementation employs a two-stage prompting strategy.

3.1.1 Prompt 1

In Prompt 1, prompts are designed to establish a comprehensive structure and define relationships within each of the five educational domains. To enhance the Claude 3.5 Sonnet model's ability to generate accurate and realistic outputs that align with real-world educational scenarios, we incorporate examples of knowledge graph structures and relationships for each domain into the prompts [16]. These exemplars, extracted from generalized educational contexts, serve as foundational scaffolding, thereby guiding the model to generate outcomes that both accurately reflect the intended pedagogical applications and maintain high degrees of precision and reliability.

1. Curriculum Development

Table 1: Example of Nodes and Relationships in KG Curriculum Development

Nodes	Course, Department, Program, LearningOutcome, Lecture
Relationships	(Course)-[:PREREQUISITE_FOR]->(Course) (Course)-[:BELONGS_TO]->(Department) (Course)-[:PART_OF]->(Program) (Course)-[:HAS_OUTCOME]->(LearningOutcome) (Lecture)-[:TEACHES]->(Course)

2. Personalized Learning

Table 2: Example of Nodes and Relationships in KG Personalized Learning

Nodes	Student, LearningStyle, LearningResource, Skill, Course
Relationships	(Student)-[:HAS_STYLE]->(LearningStyle) (Student)-[:ENROLLED_IN]->(Course) (Student)-[:HAS_SKILL]->(Skill) (LearningResource)-[:SUITABLE_FOR]->(LearningStyle) (LearningResource)-[:TEACHES]->(Skill) (Course)-[:REQUIRES]->(Skill)

3. Learning Analytics

Table 3: Example of Nodes and Relationships in KG Learning Analytics

Nodes	Student, LearningEvent, Assessment, PerformanceMetric
Relationships	(Student)-[:PARTICIPATED_IN]->(LearningEvent) (Student)-[:COMPLETED]->(Assessment) (Assessment)-[:MEASURES]->(LearningOutcome) (Student)-[:HAS_METRIC]->(PerformanceMetric)

4. Resource Management

Table 4: Example of Nodes and Relationships in KG Resource Management

Nodes	LearningResource, ResourceType, Topic, Course
Relationships	(LearningResource)-[:TYPE_OFF]->(ResourceType) (LearningResource)-[:COVERS]->(Topic) (Course)-[:USES]->(LearningResource)

5. Alumni Engagement

Table 5: Example of Nodes and Relationships in KG Alumni Engagement

Nodes	Alumni, Employer, Industry, Event, Program
Relationships	(Alumni)-[:GRADUATED_FROM]->(Program) (Alumni)-[:WORKS_FOR]->(Employer) (Employer)-[:IN_INDUSTRY]->(Industry) (Alumni)-[:ATTENDED]->(Event)

I have a knowledge graph for {5 Key Domains} with the following nodes and relationships:
{Examples of knowledge graph structures and relationships for each domain}
I want to make this knowledge graph more complex and reflective of real-world {5 Key Domains} scenarios. Could you suggest additional nodes and relationships that would enhance the graph's complexity and realism? Consider aspects such as { Nodes & Relationships} and any other relevant areas.

Figure 2: Prompt 1

3.1.2 Prompt 2

In Prompt 2, the prompt was designed to generate synthetic datasets of key domains educational scenarios. The prompt is designed by combining the output of prompt 1 which is a comprehensive structure and define relationships within each of the five key domains. and combined with the rigid rules[15] prompt to ensure Claude 3.5 Sonnet can generate synthetic datasets that are accurate and mimic real world educational scenarios of the five key dom.

I want to visualize knowledge graph in Neo4j. the dataset will be about {Key Domains}, can you generated CSV files with each containing around 50 rows tabel per synthetic datasets . Remember when u generate the synthetic data, generate dataset that mimic the real world data.
{Comprehensive Nodes & Relationships}

Figure 3: Prompt 2

3.1.3 QA (Quality Assurance)

The Quality Assurance (QA) phase marks an important checkpoint in our synthetic data generation pipeline, where human evaluators conduct a thorough assessment through three key validations. Starting with data validation, by scrutinizing each data set to ensure its accuracy and completeness, carefully examining each data point and identifying potential gaps and discrepancies. In relationship consistency validation, they verify that the relationships between educational elements make logical sense - for example, ensuring course prerequisites follow proper academic progression. Finally, educational domain alignment confirms that all data generated authentically fits the educational scenarios of 5 key domains, verifying that this synthetic data reflects real-world educational scenarios. Through this systematic validation approach, the QA phase bridges automated generation with practical educational needs, ensuring technical accuracy and educational relevance in the final knowledge graph structure. This human oversight proves essential in maintaining the quality of data linkage to educational scenarios.

3.1.4 CSV Transform

The CSV Transform phase marks the final stage in our synthetic data generation pipeline, where we utilize Python's built-in libraries csv and os - for efficient data transformation. The 'csv' library helps us handle the intricacies of CSV file operations, enabling precise control over data formatting and structure. Meanwhile, the os library manages file system operations, helping us organize and store our transformed files systematically. Together, these libraries allow us to convert our validated datasets into properly formatted CSV files, ready for bulk import into Neo4j. This straightforward yet effective Python-driven approach ensures our educational data maintains its integrity while being prepared for knowledge graph construction.

```
import csv
import os

# Fill filename
filename = 'file.csv'
# Replace with your desired folder_path
folder_path = 'csv/'  # Change 'csv' to 'css' to match your new folder name
# Check if the folder exists and create it if it does not
if not os.path.exists(folder_path):
    os.makedirs(folder_path)
# Full path to the file including the folder
full_file_path = os.path.join(folder_path, filename)
# And copy paste the data into this template
data =  """
        """
# Splitting the data into lines
lines = data.strip().split('\n')

# Parsing the data
parsed_data = [line.split(',') for line in lines]

# Writing the data to a CSV file
with open(full_file_path, 'w', newline='') as file:
    writer = csv.writer(file)
    writer.writerows(parsed_data)

print("CSV file created successfully.")
```

Figure 4: VS Code CSV Transform

3.2 Knowledge Graph Construction & Visualization

We The second phase of our methodology focuses on knowledge graph construction and visualization using Neo4j as the graph database platform. This process encompasses three interconnected stages:

- Bulk Import
- Schema Design Model
- Preview & Run

3.2.1 Bulk Import

The Bulk Import stage leverages Neo4j's capabilities for efficient large-scale data importation. Transformed CSV dataset files are imported into Neo4j using Browse Files, enabling structured data processing per key domain. This approach ensures systematic data organization while maintaining inter-domain relationship integrity.

3.2.2 Schema Design Model

In the Schema Design Model phase, we design and construct the data model for the educational knowledge graph's key domains. This comprehensive model consists of three main components: node definitions, relationship types, and property specifications. We carefully develop schema structures that reflect the educational domain's organization, focusing on how nodes connect and relate to each other. This schema design is essential for ensuring the knowledge graph accurately captures and connects various educational elements and their relationships in a meaningful way.

3.2.3 Preview & Run

The Preview & Run stage provides an interactive interface for visualizing and analysing the generated knowledge graph. This phase includes two primary components:

- Graph Visualization: Enables visual exploration of knowledge graph structures, facilitating understanding of relationships and patterns within educational data.
- Insights: Provides analytical for extracting meaningful insights from the knowledge graph.

3.3 Evaluation

Our methodology concludes with a comprehensive evaluation phase that measures the knowledge graph's effectiveness through two key metrics:

1. Performance Metrics, with measure the technical performance of the knowledge graph by analysing:
    - Graph traversal efficiency
    - Query response times
2. Data Quality Metrics, with assess the quality of knowledge representation by evaluating:
    - Relationship accuracy between educational elements
    - Data completeness across all domains
    - Alignment with educational domain requirements

3.4 Implementation Tools

Implementation conducted in this research use several tools, such as using a Pro Claude account to use Claude 3.5 Sonnet with 200k context window, using Neo4j with AuraDB

free tier, using VS code as an environment for coding, and Microsoft Excel.

4. RESULTS AND DISCUSSION

4.1 Curriculum Development

The result of Knowledge Graphs (KG) Curriculum Development using synthetic data generated by LLM shows that synthetic data can show relationships, patterns, and structures like common data in educational scenarios and provide relevant insights where KG from this Key domain is able to show plausible trends or patterns.

4.1.1 Synthetic Data Generation

The synthetic data generation process resulted in the creation of nodes and relationships that represent various entities and their interconnections. The generated data consists of distinct 16 node types and 16 relationship types that model the complex interactions in Curriculum Development key domain. The node and relationship types represent key entities including:

Table 6: LLM-generated Nodes and Relationships KG Curriculum Development

Node Types	Relationship Types
Accreditation	ACCREDITED_ BY
CareerOutcome	BELONGS_TO
Course	CONTAINS_MODULE
DegreeRequirement	GOVERNED_BY
Department	HAS_OUTCOME
IndustryPartner	HAS_REQUIREMENT
LearningOutcome	HAS_SYLLABUS
LMS	LEADS_TO
Module	OFFERS_IN
Policy	PART_OF (50)
Program	PARTNERS_WITH
ResearchProject	PREREQUISITE_FOR
Resource	SUPPORTS
Semester	USES_RESOURCE
Syllabus	USES_TOOL
Tool	UTILIZES

These nodes and relationships collectively form a comprehensive knowledge graph structure that represents the intricate connections and dependencies in Curriculum Development. and LLM successfully produced 36 dataset files of nodes and relationships. These datasets contain data synthetic data that mimic real world relationships and patterns and maintain logical consistency and meaningful relationships between entities. These dataset files ensure the knowledge graph generated is rich and diverse which can see scenario insights in curriculum development.

4.1.2 KG Construction & Visualization

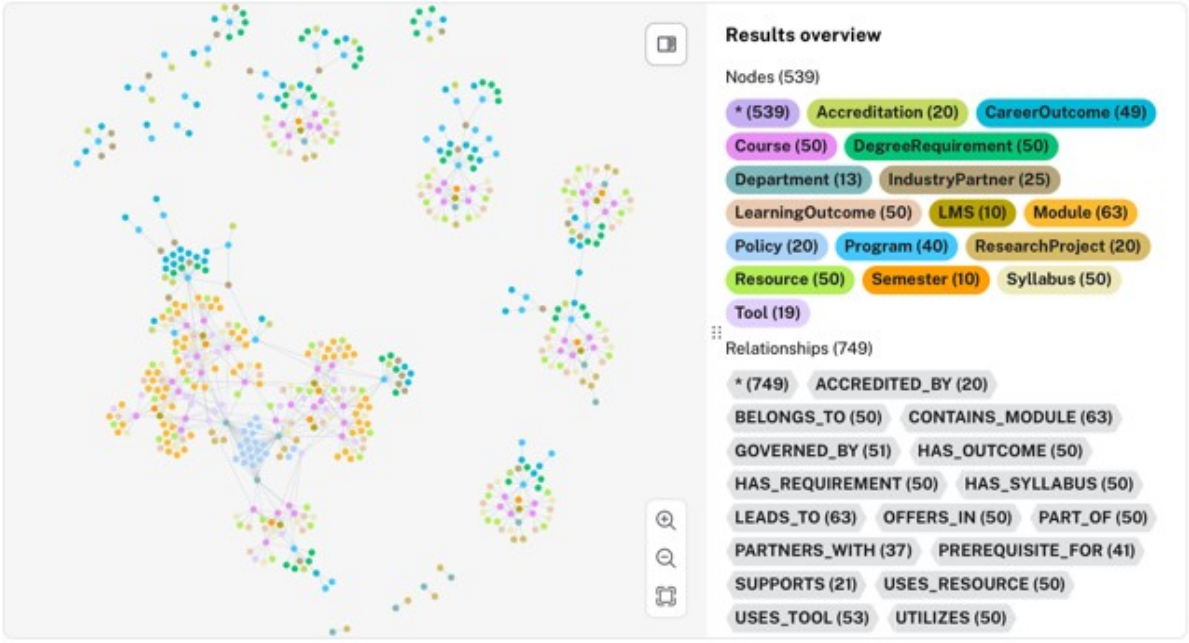

Figure 5: Knowledge Graph Curriculum Development

From KG visualization of Curriculum Development, several key insights:

1. Node Types (Primary Entities) Breakdown:
   - Academic Core
     - Course (50 nodes): Represents individual academic units

- Module (63 nodes): Largest academic component, suggesting granular content organization
- Program (40 nodes): Broader academic frameworks
- Syllabus (50 nodes): Detailed course planning documents

- Learning Elements
  - LearningOutcome (50 nodes): Specific educational objectives
  - CareerOutcome (49 nodes): Professional/career-oriented goals
  - DegreeRequirement (50 nodes): Academic completion criteria
- Support Structure
  - Resource (50 nodes): Educational materials and assets
  - Tool (19 nodes): Supporting technological/educational tools
  - LMS (10 nodes): Learning Management System components
  - Department (13 nodes): Organizational units
  - Semester (10 nodes): Temporal organization
- Quality & Partnership
  - Accreditation (20 nodes): Quality assurance elements
  - IndustryPartner (25 nodes): External professional connections
  - Policy (20 nodes): Governing guidelines
  - ResearchProject (20 nodes): Integration of research components

2. Relationship Types Analysis (749 total):

- Academic Structure Relationships
  - CONTAINS_MODULE (63): Shows modular curriculum design
  - BELONGS_TO (50): Hierarchical organization
  - PART_OF (50): Component relationships
  - OFFERS_IN (50): Course/program offerings
- Learning Path Relationships
  - LEADS_TO (63): Progressive learning pathways
  - PREREQUISITE_FOR (41): Sequential learning requirements
  - HAS_OUTCOME (50): Learning objective linkages
- Academic Structure Relationships
  - HAS_SYLLABUS (50): Course documentation
  - HAS_REQUIREMENT (50): Academic standards
  - USES_RESOURCE (50): Resource utilization
  - USES_TOOL (53): Technology integration
  - UTILIZES (50): General resource usage
- Quality & Partnership Relationships
  - ACCREDITED_BY (20): Quality assurance connections
  - PARTNERS_WITH (37): Industry collaboration
  - GOVERNED_BY (51): Policy and governance structure

3. Network Structure Analysis:

- Topology
  - Multiple interconnected clusters indicating specialized sub-domains
  - Central dense network suggesting strong cross-disciplinary integration
  - Satellite clusters possibly representing specialized programs or tracks
  - Clear hierarchical structure from programs down to modules
- Integration Patterns
  - High connectivity between academic and industry components
  - Strong linkage between learning outcomes and career outcomes
  - Comprehensive resource integration across different educational components

4. Network Structure Analysis:

- Curriculum Design
  - Modular structure enables flexible learning paths
  - Strong emphasis on outcomes-based education
  - Integration of industry requirements into academic planning
  - Clear progression pathways through prerequisites
- Quality Assurance
  - Multiple layers of governance and accreditation
  - Policy-driven framework ensuring standards
  - Regular assessment through learning outcomes
- Resource Management
  - Comprehensive tool and resource integration

- Structured LMS implementation
- Multiple support systems for learning delivery

- Resource Management
  - Comprehensive tool and resource integration
  - Structured LMS implementation
  - Multiple support systems for learning delivery
- Industry Alignment
  - Direct industry partner involvement
  - Career-oriented outcome mapping
  - Research project integration

4.1.3 Evaluation

1. Performance Metrics

- Average of 1.39 connections per node
- Time to processes nodes and relationships run import 7 seconds

2. Data Quality Metrics:

- Data Completeness: 77.55%
  - Node Retention Rate: 77.55%
  - Attribute Completeness: 77.55%
  - Shows strong preservation of essential curriculum data during transformation
- Relationship Accuracy: 45.15%
  - Relationship Density: 0.26%
  - Relationship Evenness: 75.09%
  - Reflects strategic relationship implementation focusing on meaningful connections
- Domain Alignment: 82.84%
  - Curriculum domain: 63.25%
  - Technical domain: 98.75%
  - Administrative domain: 90.00%
  - Industry domain: 98.95%
  - Weighted based on domain importance (Curriculum: 40%, Others: 20% each)

3. Key Insights:

- The Relationship Accuracy score (45.15%) indicates a deliberate approach to relationship creation, prioritizing meaningful pedagogical connections over comprehensive connectivity, which is beneficial for curriculum management and navigation.
- The Data Completeness score (77.55%) demonstrates effective preservation of critical curriculum data while successfully removing redundant or obsolete elements.
- The Domain Alignment score (82.84%) shows exceptionally strong preservation of industry and technical domains, while maintaining solid administrative functionality. The lower curriculum domain score reflects intentional optimization of module structure.
- The overall average of 68.52% reflects a successful transformation that achieved its primary goals of streamlining the curriculum structure while maintaining critical relationships and domain integrity. The high scores in technical and industry domains (98.75% and 98.95%) particularly highlight the successful preservation of key external partnerships and technical infrastructure.

4.2 Personalized Learning

The result of Knowledge Graphs (KG) Personalized Learning using synthetic data generated by LLM shows that synthetic data can show relationships, patterns, and structures like common data in educational scenarios and provide relevant insights where KG from this Key domain is able to show plausible trends or patterns.

4.2.1 Synthetic Data Generation

The synthetic data generation process resulted in the creation of nodes and relationships that represent various entities and their interconnections. The generated data consists of distinct 6 node types and 5 relationship types that model the complex interactions in Personalized Learning key domain. The node and relationship types represent key entities including:

Table 7: LLM-generated Nodes and Relationships KG Personalized Learning

Node Types	Relationship Types
Course	ENROLLED_IN
Learning Style	HAS_STYLE
Learning Resource	REQUIRES
Skill	SUITABLE_FOR
Student	TEACHES
Relationships	

These nodes and relationships collectively form a comprehensive knowledge graph structure that represents the intricate connections and dependencies in Personalized Learning. and LLM successfully produced 11 dataset files of nodes and relationships. These datasets contain data synthetic data that mimic real world relationships and patterns and maintain logical consistency and meaningful relationships between entities. These

Journal of Theoretical and Applied Information Technology
28th February 2026. Vol.104. No.4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

dataset files ensure the knowledge graph generated is rich and diverse which can see scenario insights in Personalized Learning.

4.2.2 KG Construction & Visualization

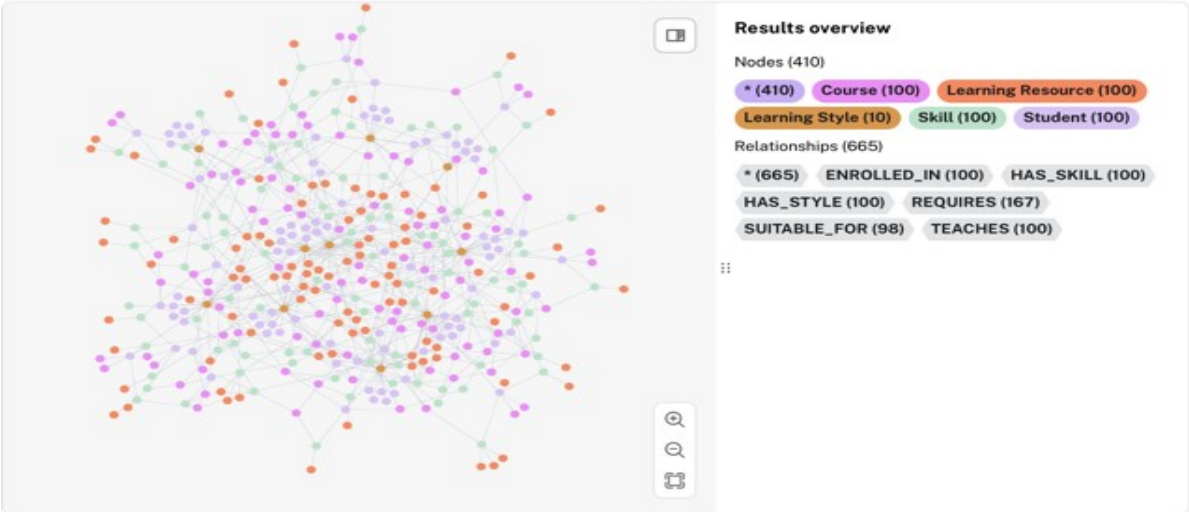

Figure 6: Knowledge Graph Personalized Learning

From KG visualization of Personalized Learning, several key insights:

1. KG Overview:
- Total Nodes: 410
- Total Relationships: 665
- More centralized structure, suggesting tight integration between components

2. Node Types (Primary Entities) Breakdown:
- Core Educational Components
  - Course (100 nodes): Core instructional units
  - Learning Resource (100 nodes): Educational materials and content
  - Student (100 nodes): Individual learners
  - Skill (100 nodes): Specific competencies
  - Learning Style (10 nodes): Different approaches to learning

3. Relationship Types Analysis (665 total)
- Student-Centered Relationships
  - ENROLLED_IN (100): Student-course connections
  - HAS_SKILL (100): Student competency mapping
  - HAS_STYLE (100): Learning style preferences
- Course-Related Relationships
  - REQUIRES (167): Highest number of relationships, indicating prerequisite skills/knowledge
  - TEACHES (100): Learning outcomes and skill development
  - SUITABLE_FOR (98): Learning style compatibility

4. Key Insights
- Personalization Focus
  - Strong emphasis on learning styles (though smaller number suggests consolidated categories)
  - Direct mapping between students, skills, and learning styles
  - Resource suitability based on learning preferences
- Skill-Based Architecture
  - Equal distribution of courses and skills (100 each)
  - Heavy emphasis on skill requirements (167 REQUIRES relationships)
  - Clear skill development pathways through courses
- Resource Integration
  - One-to-one ratio of courses to learning resources
  - Resources likely tailored to different learning styles
  - Structured alignment with skill development

5. Educational Design Implications
- Personalized Learning

- System supports multiple learning pathways
- Accommodates different learning styles
- Resources matched to individual preferences

- Competency-Based Education
  - Strong focus on skill acquisition
  - Clear prerequisite structure
  - Direct mapping of skills to courses
- Student-Centered Approach
  - Individual learning style recognition
  - Personalized resource allocation
  - Skill-based progression tracking

6. Structural Analysis

- Network Characteristics
  - Dense central clustering indicates high interconnectivity
  - Peripheral nodes suggesting specialized or advanced topics
  - Balanced distribution of different node types
- Integration Patterns
  - Strong course-skill-resource triangulation
  - Learning style considerations across components
  - Student progression pathways clearly visible

This knowledge graph represents a modern, learner-centric educational system with:

- Strong personalization capabilities
- Clear skill development pathways
- Integrated resource management
- Flexible learning approaches

The structure supports adaptive learning while maintaining clear progression paths and skill development objectives. The balanced distribution of nodes suggests a well-planned system designed to accommodate diverse learning needs while ensuring consistent educational outcomes.

4.2.3 Evaluation

1. Performance Metrics

- Average of 1.62 connections per node
- Time to processes nodes and relationships run import 4 seconds

2. Data Quality Metrics:

- Data Completeness: 100%
  - Node Retention Rate: 100.00%
  - Attribute Completeness: 100.00%
  - Perfect preservation of all node types and attributes in the personalized learning system
- Relationship Accuracy: 46.55%
  - Relationship Density: 0.40%
  - Relationship Evenness: 77.33%
  - Strategic implementation of personalized learning relationships with balanced distribution
- Domain Alignment: 100.00%
  - Learner domain: 100.00%
  - Content domain: 100.00%
  - Competency domain: 100.00%
  - Weighted based on domain importance (Content: 40%, Learner: 30%, Competency: 30%)

3. Key Insights:

- The Relationship Accuracy score (46.55%) reflects a highly targeted approach to creating learning relationships, focusing on meaningful connections between students, courses, and skills rather than exhaustive connectivity. The high evenness score (77.33%) indicates well-balanced relationship distribution across different types of learning interactions.
- The perfect Data Completeness score (100.00%) demonstrates full preservation of all learning entities and their attributes, ensuring no loss of critical educational data during the transformation. This is particularly important for maintaining comprehensive learner profiles and educational resources.
- The Domain Alignment score (100.00%) shows perfect maintenance across all three crucial domains (learner, content, and competency), with appropriate weighting that prioritizes content delivery while balancing learner needs and skill development.
- The overall average of 82.18% indicates a highly successful transformation that achieved its primary goal of creating a personalized learning environment while maintaining complete data integrity. The knowledge graph effectively captures the complex relationships between students, learning styles, courses, resources, and skills, enabling sophisticated personalization capabilities.

4.3 Learning Analytics

The result of Knowledge Graphs (KG) Learning Analytics using synthetic data generated by LLM shows that synthetic data can show relationships, patterns, and structures like common data in educational scenarios and provide relevant insights where KG from this Key domain is able to show plausible trends or patterns.

4.3.1 Synthetic Data Generation

The synthetic data generation process resulted in the creation of nodes and relationships that represent various entities and their interconnections. The generated data consists of distinct 5 node types and 4 relationship types that model the complex interactions in Learning Analytics key domain. The node and relationship types represent key entities including:

Table 8: LLM-generated Nodes and Relationships KG Learning Analytics

Node Types	Relationship Types
Assessment	COMPLETED
Learning Event	HAS_METRIC
Learning Outcome	MEASURES
PerformanceMetric	PARTICIPATED_IN
Student	

These nodes and relationships collectively form a comprehensive knowledge graph structure that represents the intricate connections and dependencies in Learning Analytics. and LLM successfully produced 9 dataset files of nodes and relationships. These datasets contain data synthetic data that mimic real world relationships and patterns and maintain logical consistency and meaningful relationships between entities. These dataset files ensure the knowledge graph generated is rich and diverse which can see scenario insights in Learning Analytics.

4.3.2 KG Construction & Visualization

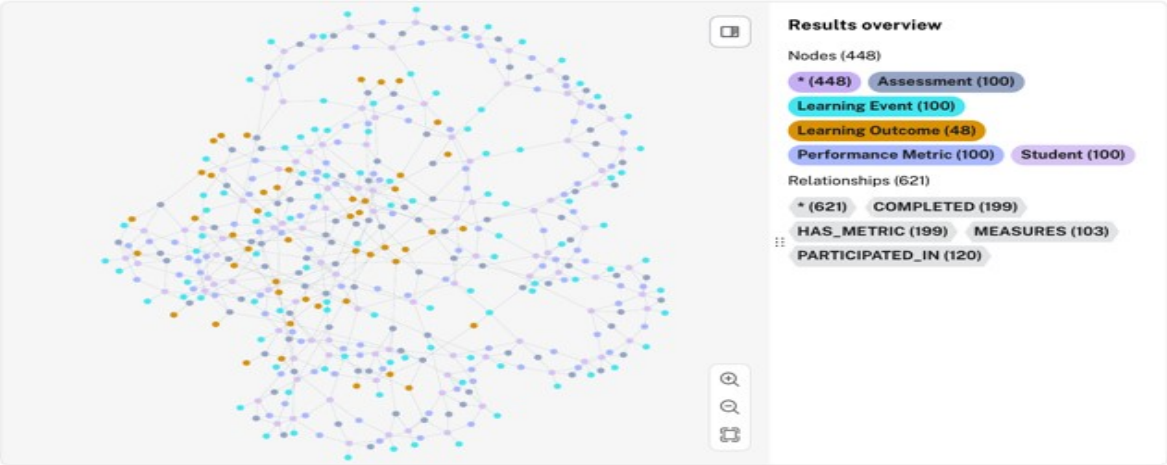

Figure 7: Knowledge Graph Learning Analytics

From KG visualization of Learning Analytics, several key insights:

1. KG Overview:
   - Nodes (448 total)
     - Assessment (100 nodes)
     - Learning Event (100 nodes)
     - Learning Outcome (48 nodes)
     - Performance Metric (100 nodes)
     - Student (100 nodes)
   - Relationships (621 total)
     - COMPLETED (199 connections)
     - HAS_METRIC (199 connections)
     - MEASURES (103 connections)
     - PARTICIPATED_IN(120 connections)
2. Key Insights:
   - Comprehensive Data Structure
     - The graph shows a well-connected educational ecosystem tracking multiple aspects of the learning process
     - There's an equal distribution (100 each) of Assessments, Learning Events, Performance Metrics, and Students, suggesting systematic data collection
   - Learning Outcome Focus
     - With 48 Learning Outcomes connected to other nodes, the system appears to be outcomes-based

- Each Learning Outcome likely maps to multiple Assessments and Learning Events
- Student Engagement Patterns
  - The "PARTICIPATED_IN" relationship (120) shows student participation in learning events
  - "COMPLETED" relationships (199) likely track assessment completion
- Structure Assessment & Measurement
  - Strong connection between Assessments and Performance Metrics (HAS_METRIC: 199)
  - The "MEASURES" relationship (103) suggests systematic evaluation of learning outcomes
- Network Density
  - The visualization shows a dense, interconnected network
  - Multiple pathways between nodes suggest rich data relationships for tracking student progress

3. Practical Applications:

- Student Progress Tracking, the system can track individual student journeys through learning events and assessments
- Outcome Achievement, can measure how effectively learning events and assessments align with intended outcomes
- Performance Analysis, rich metric relationships enable detailed performance analysis
- Program Evaluation, the interconnected nature allows evaluation of entire educational programs.

4.3.3 Evaluation

1. Performance Metrics

- Average of 1.39 connections per node
- Time to processes nodes and relationships run import 3 seconds

2. Data Quality Metrics:

- Data Completeness: 99.56%
  - Node Retention Rate: 99.56%
  - Attribute Completeness: 99.56%
  - Near-perfect preservation of analytics entities with minimal optimization.
- Relationship Accuracy: 43.06%
  - Relationship Density: 0.31%
  - Relationship Evenness: 71.55%
  - Strategic implementation of analytics-focused relationship types
- Domain Alignment: 99.60%
  - Performance domain: 100.00%
  - Learning domain: 98.67%
  - Student domain: 100.00%
  - Weighted based on domain importance (Performance: 40%, Learning: 30%, Student: 30%)

3. Key Insights:

- The Relationship Accuracy score (43.06%) demonstrates a highly focused approach to analytics relationships, emphasizing key performance tracking connections (COMPLETED, HAS_METRIC) and learning event participation (PARTICIPATED_IN, MEASURES). The high evenness score (71.55%) indicates well-balanced distribution across these critical measurement relationships.
- The exceptional Data Completeness score (99.56%) reflects near-perfect preservation of all analytics entities, with only minimal optimization in learning outcomes (50 to 48 nodes). This high retention ensures comprehensive analytics capabilities while removing potential redundancies.
- The strong Domain Alignment score (99.60%) shows excellent maintenance of domain integrity across all three core areas: Perfect retention in performance tracking (100%), Near-perfect preservation of learning analytics (98.67%), and Complete maintenance of student tracking capabilities (100%)
- The overall average of 80.74% indicates a highly successful transformation that achieved its primary goals: Establishing clear performance tracking pathways, Maintaining comprehensive learning analytics coverage, Preserving student-centered measurement capabilities, and Creating efficient, meaningful relationships between analytics entities.

4.4 Resource Management

The result of Knowledge Graphs (KG) Resource Management using synthetic data generated by LLM shows that synthetic data can show relationships, patterns, and structures like common data in educational scenarios and provide relevant insights where KG from this Key domain is able to show plausible trends or patterns.

4.4.1 Synthetic Data Generation

The synthetic data generation process resulted in the creation of nodes and relationships that represent various entities and their interconnections. The generated data consists of distinct 14 node types and 14 relationship types that model the complex interactions in Resource Management key domain. The node and relationship types represent key entities including:

Table 9: LLM-generated Nodes and Relationships KG Resource Management

Node Types	Relationship Types
Author	AVAILABILITY_
Availability	STATUS
Course	COVERS
Format	CREATES
DigitalRepository	ENROLLED_IN
Instructor	IN_FORMAT
Language	IN_LANGUAGE
LearningResource	LICENSED_UNDER
License	PROVIDES
Publisher	PUBLISHED_BY
ResourceType	REVIEWS
Review	STORED_IN
Topic	TEACHES
User	TYPE_OF
	USES

These nodes and relationships collectively form a comprehensive knowledge graph structure that represents the intricate connections and dependencies in Learning Analytics. and LLM successfully produced 29 dataset files of nodes and relationships. These datasets contain data synthetic data that mimic real world relationships and patterns and maintain logical consistency and meaningful relationships between entities. These dataset files ensure the knowledge graph generated is rich and diverse which can see scenario insights in Learning Analytics.

4.4.2 KG Construction & Visualization

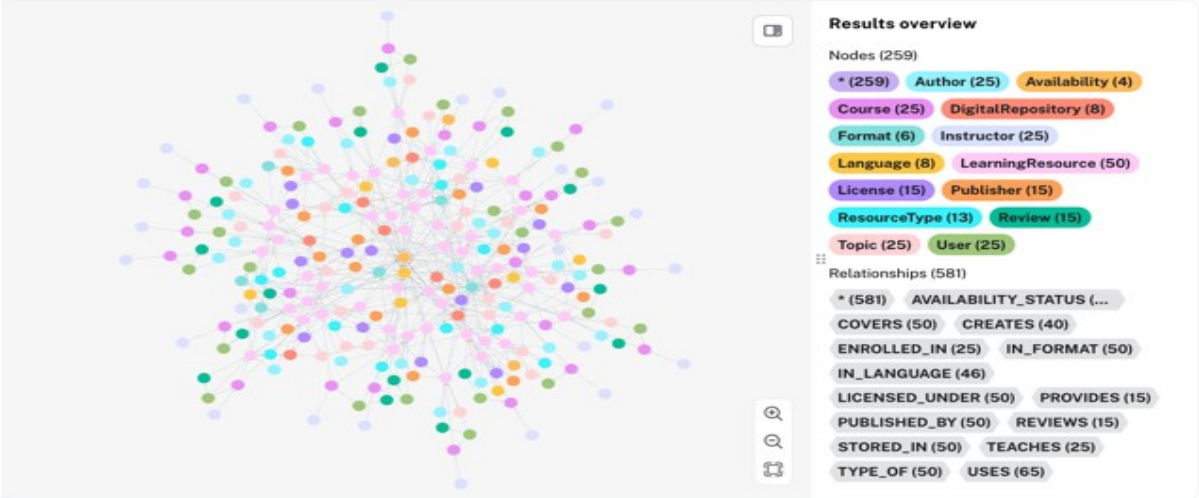

Figure 8: Knowledge Graph Resource Management

From KG visualization of Resource Management, several key insights:

1. KG Overview:
   - Nodes (259 total)
     - Author (25 nodes)
     - Course (25 nodes)
     - DigitalRepository (8 nodes)
     - Format (6 nodes)
     - Instructor (25 nodes)
     - Language (8 nodes)
     - LearningResource (50 nodes)
     - License (15 nodes)
     - Publisher (15 nodes)
     - ResourceType (13 nodes)
     - Review (15 nodes)
     - Topic (25 nodes)
     - User (25 nodes)
     - Availability (4 nodes)
   - Relationships (581 total)
     - COVERS (50 connections)
     - CREATES (40 connections)
     - ENROLLED_IN (25 connections)
     - IN_FORMAT (50 connections)
     - IN_LANGUAGE (46 connections)
     - LICENSED_UNDER (50 connections)
     - PROVIDES (15 connections)
     - PUBLISHED_BY (50 connections)
     - REVIEWS (15 connections)
     - STORED_IN (50 connections)
     - TEACHES (25 connections)

- USES (65 connections)
- AVAILABILITY_STATUS

2. Key Insights:

- Resource-Centric Structure
  - LearningResource (50 nodes) forms the core of the system
  - Strong metadata management with multiple descriptive attributes (format, language, license, type)
  - Comprehensive tracking of resource lifecycle from creation to usage
- Content Management
  - DigitalRepositories managing the storage
  - Multiple formats (6) and resource types (13) indicating diverse content
  - Robust licensing system (15 different licenses)
- User Ecosystem
  - Balanced distribution between Authors (25), Instructors (25), and Users (25)
  - Clear distinction between content creators and consumers
  - Multiple roles in the content lifecycle
- Access and Availability
  - different availability statuses suggest structured access control
  - Multiple languages (8) indicating international/multilingual support
  - Various formats supporting different access methods
- Quality Control
  - Review system in place (15 review nodes)
  - Publisher involvement (15 publishers) suggesting quality standards
  - Licensed content management (50 LICENSED_UNDER relationships)

3. Practical Applications:

- Resource Discovery
  - Multi-dimensional search capabilities (by topic, language, format)
  - Clear content categorization
  - Easy tracking of resource availability
- Content Management
  - Structured storage across digital repositories
  - Clear licensing and publishing workflows
  - Format and type management for different use cases
- Personalization Focus
  - Course-resource alignment
  - Instructor-content relationships
  - Student enrolment tracking

4.4.3 Evaluation

4. Performance Metrics:

- Average of 2.24 connections per node
- Time to processes nodes and relationships run import 8 seconds

5. Data Quality Metrics:

- Data Completeness: 89.93%
  - Node Retention Rate: 89.93%
  - Attribute Completeness: 89.93%
  - High score indicating successful data preservation during transformation
- Relationship Accuracy: 38.21%
  - Relationship Density: 0.87%
  - Relationship Evenness: 63.11%
  - Lower score due to selective relationship implementation rather than full connectivity
- Domain Alignment: 89.89%
  - Educational domain: 100%
  - Technical domain: 60%
  - Administrative domain: 89.47%
  - User domain: 100%
  - Weighted based on domain importance (Educational: 40%, Others: 20% each)

6. Key Insights:

- The relatively low Relationship Accuracy score (38.21%) suggests a selective and purposeful approach to relationship creation rather than a fully connected graph, which is actually desirable for maintainability and performance.
- The high Data Completeness score (89.93%) indicates excellent preservation of critical data during the knowledge graph transformation.
- The strong Domain Alignment score (89.89%) shows successful preservation of domain integrity, particularly in educational and user-centered aspects.
- The overall average of 72.68% reflects a successful transformation that prioritized meaningful connections and domain preservation while optimizing technical infrastructure.

Journal of Theoretical and Applied Information Technology
28th February 2026. Vol.104. No.4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.5 Alumni Engagement

The result of Knowledge Graphs (KG) Alumni Engagement using synthetic data generated by LLM shows that synthetic data can show relationships, patterns, and structures like common data in educational scenarios and provide relevant insights where KG from this Key domain is able to show plausible trends or patterns.

4.5.1 Synthetic Data Generation

The synthetic data generation process resulted in the creation of nodes and relationships that represent various entities and their interconnections. The generated data consists of distinct 10 node types and 12 relationship types that model the complex interactions in Alumni Engagement key domain. The node and relationship types represent key entities including:

Table 10: LLM-generated Nodes and Relationships KG Alumni Engagement

Node Types	Relationship Types
Alumni CareerDevelopment-Program Employer Event NetworkingGroup Scholarship	ACTIVE_ON ATTENDED DONATED FUNDRAISES_FOR FUNDS GRADUATED_FROM HOSTS IN_INDUSTRY MEMBER_OF PARTICIPATES_IN RECEIVED WORKS_FOR

These nodes and relationships collectively form a comprehensive knowledge graph structure that represents the intricate connections and dependencies in Alumni Engagement. and LLM successfully produced 11 dataset files of nodes and relationships. These datasets contain data synthetic data that mimic real world relationships and patterns and maintain logical consistency and meaningful relationships between entities. These dataset files ensure the knowledge graph generated is rich and diverse which can see scenario insights in Alumni Engagement.

4.5.2 KG Construction & Visualization

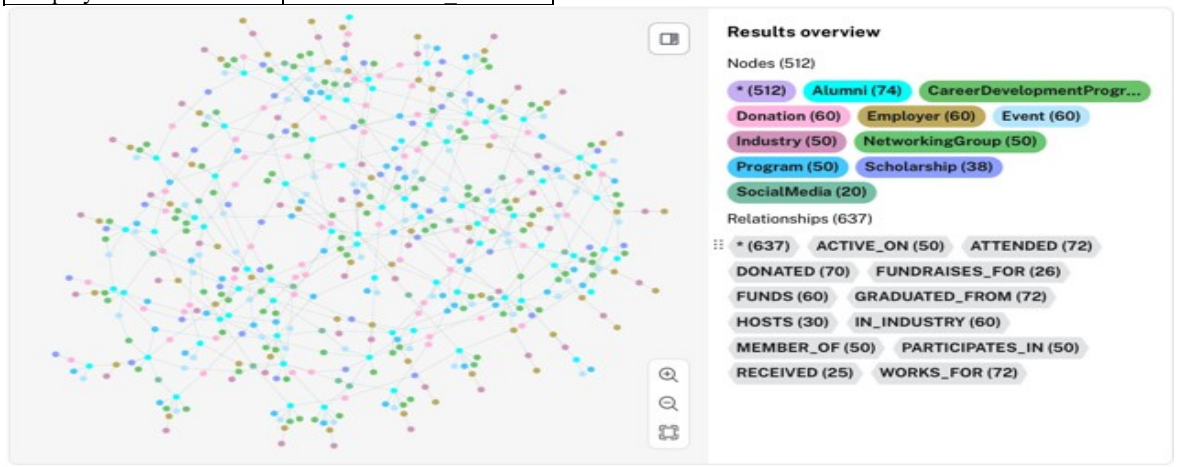

Figure 9: Knowledge Graph Alumni Engagement

From KG visualization of Alumni Engagement, several key insights:

1. KG Overview:
   - Nodes (512 total)
     - Alumni (74 nodes)
     - CareerDevelopmentProgram(50 nodes)
     - Donation (60 nodes)
     - Employer (60 nodes)
     - Event (60 nodes)
     - Industry (50 nodes)
     - NetworkingGroup (50 nodes)
     - Program (50 nodes)
     - Scholarship (38 nodes)
     - SocialMedia (20 nodes)
   - Relationships (637 total)
     - ACTIVE_ON (50 connections)
     - ATTENDED (72 connections)
     - DONATED (70 connections)
     - FUNDRAISES_FOR (26 connections)
     - FUNDS (60 connections)
     - GRADUATED_FROM (72 connections)
     - HOSTS (30 connections)
     - IN_INDUSTRY (60 connections)
     - MEMBER_OF (50 connections)

- PARTICIPATES_IN (50 connections)
- RECEIVED (25 connections)
- WORKS_FOR (72 connections)

2. Key Insights:

- Alumni Engagement
  - Strong focus on alumni involvement (74 alumni nodes)
  - Multiple engagement channels (events, networking groups, social media)
  - Active participation tracking through various relationships
- Career Development
  - Robust employer relationships (60 employers)
  - Industry diversity (50 different industries)
  - Career program integration
  - Strong employment tracking (72 WORKS_FOR relationships)
- Fundraising and Giving
  - Significant donation activity (60 donation nodes)
  - Scholarship program management (38 scholarships)
  - Multiple fundraising pathways (26 FUNDRAISES_FOR relationships)
  - Active donor base (70 DONATED relationships)
- Networking Infrastructure
  - Dedicated networking groups (50)
  - Event-based networking (60 events)
  - Social media integration (20 platforms/channels)
  - Strong membership tracking (50 MEMBER_OF relationships)

3. Practical Applications:

- Alumni Relations Management
  - Track alumni career progression
  - Monitor engagement levels
  - Manage giving relationships
  - Facilitate networking opportunities
- Career Services
  - Connect current students with alumni
  - Track industry placement trends
  - Manage employer relationships
  - Coordinate career development programs
- Fundraising Operations
  - Monitor donation patterns
  - Track scholarship distribution
  - Identify potential donors
  - Manage fundraising events
- Event Management
  - Coordinate networking events
  - Track attendance patterns
  - Measure engagement success
  - Optimize event planning

4.5.3 Evaluation

1. Performance Metrics

- Average of 1.24 connections per node
- Time to processes nodes and relationships run import 6 seconds

2. Data Quality Metrics:

- Data Completeness: 97.52%
  - Node Retention Rate: 97.52%
  - Attribute Completeness: 97.52%
  - High retention with strategic optimization in scholarship management
- Relationship Accuracy: 40.69%
  - Relationship Density: 0.24%
  - Relationship Evenness: 67.65%
  - Diverse relationship types covering engagement, career, and support activities
- Domain Alignment: 97.55%
  - Engagement domain: 99.51%
  - Career domain: 100.00%
  - Support domain: 92.50%
  - Weighted based on domain importance (Engagement: 40%, Career: 30%, Support: 30%)

3. Key Insights:

- The Relationship Accuracy score (40.69%) reflects a strategic implementation of alumni engagement connections, with well-distributed relationship types across: Event participation (ATTENDED: 72), Career tracking (WORKS_FOR: 72, IN_INDUSTRY: 60), Giving patterns (DONATED: 70, FUNDS: 60), and Community involvement (MEMBER_OF: 50, PARTICIPATES_IN: 50).
- The strong Data Completeness score (97.52%) demonstrates excellent preservation of alumni network data, with minimal and strategic reductions: One alumni record optimization (75 to 74), Targeted scholarship program streamlining (50 to 38), and perfect

retention in career development and industry connections.

- The Domain Alignment score (97.55%) shows exceptional maintenance of core functions: Near-perfect engagement preservation (99.51%), Complete career services retention (100%), and strategic optimization of support services (92.50%)
- The overall average of 78.59% indicates a successful transformation that: Enhanced alumni networking capabilities, maintained strong career development connections, Optimized support programs for efficiency, and created meaningful relationship pathways across all engagement dimensions.

4.6 Discussion

The research findings provide compelling evidence for the effectiveness of LLM-generated synthetic data in developing educational knowledge graphs. Through systematic implementation across five educational domains, our study demonstrates the capability of this approach to accurately model complex educational relationships while offering practical benefits in terms of development efficiency and testing flexibility. The ability of LLM-generated synthetic data to mimic real-world relationships and patterns emerged as a significant strength of our approach. In the curriculum development domain, the synthetic data successfully captured intricate academic hierarchies and dependencies through a complex network of 16 node types and 16 relationship types. This structural complexity was matched by remarkably high domain alignment scores, particularly in personalized learning (100%) and learning analytics (99.60%), indicating that the generated data accurately reflects the multifaceted nature of educational relationships. The relationship accuracy metrics, while seemingly modest at 38.21-46.55% across domains, actually represent a deliberate focus on meaningful connections that mirror real educational scenarios rather than exhaustive but superficial linkages.

The validity of insights generated through this approach is evidenced by the comprehensive analytical capabilities demonstrated across all domains. In curriculum development, the knowledge graphs revealed clear visualization of prerequisite chains and learning outcome alignments, providing valuable tools for curriculum planning and assessment. The personalized learning domain showed particularly strong results in mapping individual learning pathways and resource recommendations, while the learning analytics domain demonstrated robust capabilities in tracking student progress and performance patterns. Resource management graphs effectively captured resource lifecycle patterns, and alumni engagement successfully mapped complex professional networks and career trajectories. The consistency in data quality metrics across domains (77.55-100% completeness) further supports the reliability of these insights for educational decision-making. One of the most significant advantages of our approach lies in its efficiency regarding cost and time resources. The two-stage prompting strategy enabled rapid synthetic data generation, with processing times averaging just 3-8 seconds per domain. This represents a dramatic reduction compared to traditional data collection methods, which often require months of gathering and cleaning data. Furthermore, the implementation proved remarkably resource-efficient, operating effectively within the constraints of Neo4j's free tier and requiring minimal computational resources. This efficiency extends to human resource requirements as well, with streamlined quality assurance processes and reduced dependency on extensive data collection teams. The framework's capability for scenario testing emerged as another crucial strength. The synthetic data approach demonstrates remarkable flexibility in accommodating various educational scenarios, relationship configurations, and scale simulations. This adaptability proves particularly valuable for educational institutions seeking to test different approaches to curriculum design, student assessment, resource allocation, or alumni engagement strategies. The ease with which parameters can be modified allows for rapid iteration and refinement of educational models, supporting evidence-based decision-making in educational planning and administration.

4.7 Limitations

Despite the promising results, several important limitations warrant consideration in evaluating this research. The most significant challenge lies in the validation of synthetic data authenticity. While our metrics show high structural accuracy and domain alignment, the absence of direct comparison with real educational datasets leaves some uncertainty about the completeness of our relationship patterns. Complex educational interactions, particularly

those that evolve over time, may not be fully captured in our current implementation. The validation constraints extend beyond data authenticity to practical application. Without extensive feedback from educational stakeholders or longitudinal studies in real educational settings, the full practical utility of our approach remains to be confirmed. This limitation is particularly relevant when considering the nuanced requirements of different educational institutions and the varying complexity of their data needs.

Technical limitations, primarily stemming from our use of Neo4j's free tier, restricted our ability to fully test the system's scalability and performance under heavy loads. While the current implementation demonstrates efficient processing for our test cases, questions remain about performance optimization for larger-scale deployments and more complex query requirements. Methodological constraints also merit consideration. The current approach relies heavily on predefined relationship structures, potentially limiting its ability to discover novel or unexpected patterns in educational data. The quality assurance process, while systematic, introduces an element of subjectivity that could influence the final knowledge graph structure. Additionally, the exploration of cross-domain relationships remains somewhat limited, potentially missing important interconnections between different educational aspects. Implementation challenges present another set of limitations. The effectiveness of the system depends significantly on careful prompt engineering and requires manual intervention in the quality assurance process. These requirements, combined with the need for expertise in both educational domain knowledge and technical implementation, may present barriers to widespread adoption.

5. CONCLUSION

This research provides substantial evidence that LLM-generated synthetic data can effectively serve as a prototype for developing educational knowledge graphs that accurately reflect real-world relationships and patterns. The high levels of structural accuracy (data completeness: 77.55-100%) and domain alignment (82.84-100%) achieved across five educational domains demonstrate the viability of this approach for modelling complex educational relationships while addressing critical concerns about data privacy and accessibility. The findings validate the effectiveness of our approach through multiple dimensions. The synthetic data successfully mimicked complex educational relationships and patterns, generating valid, actionable insights across all studied domains. The significant reduction in data collection time and costs, combined with flexible scenario testing capabilities, suggests that this approach could substantially improve the efficiency of educational data modelling and analysis.

However, the identified limitations point to important areas for future research. These include the need for validation against real educational datasets, expansion of scenario testing capabilities, development of more sophisticated validation methods, and comprehensive integration testing with existing educational systems. These challenges, while significant, do not diminish the fundamental value demonstrated by our approach. Our research makes a significant contribution to educational data management by establishing a viable methodology for developing knowledge graphs using synthetic data. This approach provides institutions with a cost-effective method for prototyping and testing data-driven solutions while maintaining privacy and security. The findings suggest that LLM-generated synthetic data can effectively bridge the gap between the need for comprehensive educational data modelling and the practical constraints of real-world data collection, offering a promising path forward for educational institutions seeking to leverage data-driven insights while protecting sensitive information.

ACKNOWLEDGMENT

Author Contributors: Tanty Oktavia - Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Data Collection, Visualization, Writing, and Editing; M Bagaskoro Triwicaksana - Conceptualization, Formal Analysis, Project Administration, Validation, and Review. The data supporting this study have been made publicly available to facilitate further research. The dataset can be accessed through Zenodo under the DOI: https://doi.org/10.5281/zenodo.18161406.

REFERENCES

[1] B. Abu-Salih and S. Alotaibi, “A systematic literature review of knowledge graph construction and application in education,” Feb. 15, 2024, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2024.e25383.

[2] G. Agrawal, Y. Deng, J. Park, H. Liu, and Y. C. Chen, "Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cybersecurity Education," *Information (Switzerland)*, vol. 13, no. 11, Nov. 2022, doi: 10.3390/info13110526.

[3] N. Hubert, A. Brun, and D. Monticolo, "New Ontology and Knowledge Graph for University Curriculum Recommendation," 2022. [Online]. Available: http://ceur-ws.org

[4] Z. Li, L. Cheng, C. Zhang, X. Zhu, and H. Zhao, "Multi-source Education Knowledge Graph Construction and Fusion for College Curricula," May 2023, doi: 10.1109/ICALT58122.2023.00111.

[5] J. Barrasa and J. Webber, "Building Knowledge Graphs," 2022.

[6] Y. Tang, E. Yuan, W. Chen, S. Zhang, L. Liu, and Y. Wu, "Analysis of Learning Effectiveness Based on Knowledge Graph," 2023, pp. 1742–1749. doi: 10.2991/978-94-6463-172-2_193.

[7] Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations."

[8] V. C. Pezoulas *et al.*, "Synthetic data generation methods in healthcare: A review on open-source tools and methods," Dec. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.csbj.2024.07.005.

[9] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.05229

[10] B. Zhao *et al.*, "EDUKG: a Heterogeneous Sustainable K-12 Educational Knowledge Graph," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.12228

[11] Y. Su and Y. Zhang, "Automatic Construction of Subject Knowledge Graph based on Educational Big Data," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Apr. 2020, pp. 30–36. doi: 10.1145/3396452.3396458.

[12] K. Wiggers and D. Coldewey, "This Week in AI: Tech giants embrace synthetic data," TechCunch.

[13] Y. Liu, Y. Yin, W. Cheng, and C. Li, "Construction and Application of User Check-in Spatiotemporal Knowledge Graph Based on Neo4j," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 609–616. doi: 10.1016/j.procs.2024.08.117.

[14] J. Benton *et al.*, "Sabotage Evaluations for Frontier Models," Oct. 2024.

[15] OpenAI, "Prompt engineering," openai.com. Accessed: Apr. 01, 2024. [Online]. Available: https://platform.openai.com/docs/guides/prompt-engineering

[16] R. Eldan and Y. Li, "TinyStories: How Small Can Language Models Be and Still Speak Coherent English?," 2023.