Optimizing Automated Question Generation for Educational Assessments

A Semantic Analysis of LLMs with Structured and Unstructured Ontologies

Sumayyah Alamoudi
Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia
ssalehalamoudi0001@stu.kau.edu.sa (corresponding author)

Lama A. Al Khuzayem
Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia
lalkhuzayem@kau.edu.sa

Amani Jamal
Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia
atjamal@kau.edu.sa

Received: 20 February 2025 | Revised: 22 March 2025 | Accepted: 2 April 2025

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.10662

ABSTRACT

This study explores the optimization of Automated Question Generation (AQG) for educational assessments using Large Language Models (LLMs) and ontologies. Three approaches are evaluated: template-based structured ontology question generation, LLM-based structured ontology question generation, and LLM-based flat concept list question generation, using BERT Precision, Recall, F1-score, and Semantic Similarity as performance metrics. The results show that: i) the template-based structured ontology approach achieved a BERT Precision of 0.833, Recall of 0.844, and F1-score of 0.838, with a Semantic Similarity of 0.563, ii) the LLM-based structured ontology method showed improvements with a BERT Precision of 0.856, Recall of 0.863, and F1-score of 0.859, but a lower Semantic Similarity of 0.534, and iii) the LLM-based flat concept list approach provided the best results, achieving BERT Precision, Recall, and F1-score of 0.859, along with the highest Semantic Similarity of 0.567. Despite the higher semantic similarity of the LLM-based flat concept list, qualitative analysis revealed that the unstructured ontology sometimes produced hallucinated or unrelated questions. These findings suggest that LLM-based methods provide a balance of relevance and diversity in question generation, with LLM-based flat concept list offering the most optimal results for question generation, while LLM-based structured ontology strikes a balance between Precision and Recall.

Keywords-AI in education; ontologies; question generation; Large Language Models (LLMs)

I. INTRODUCTION

AQG can potentially transform the field of educational assessments by considerably reducing teachers' workloads, streamlining the creation of evaluation tools, and enabling the personalization of learning opportunities to better address each student's specific needs and abilities. However, current AQG methods face considerable limitations in producing comprehensive and meaningful assessments, as they often struggle to generate diverse questions aligned with specific learning objectives [1].

With the growing interest in LLMs, several studies have explored their application in educational question generation [2]. A comprehensive review in [3] discussed the application of ontologies in education, where they serve as structured frameworks defining concept relationships, thereby improving knowledge representation and management. An architecture integrating ontologies with question-generation algorithms was proposed in [4]. Their integration addresses the contextual limitations of LLMs and the static nature of traditional ontology-based systems, leading to improved question quality and better alignment with instructional goals [5].

Similarly, an ontology-based Multiple-Choice Question (MCQ) generator demonstrated that structured information facilitates semantically coherent question generation [6], with further studies exploring direct MCQ generation from ontologies [7, 8]. Instructional question generation leveraging LLMs' human-like capacities was explored in [9], while authors in [10] integrated ontologies with Retrieval-Augmented Generation (RAG) to enhance LLM-generated content. Additionally, LexExMachinaQA was introduced to automatically generate ontology lexicons for linked data question answering [11]. GPT-3.5 Turbo has also been applied to generate cloze questions for English vocabulary assessments [12–14]. Other studies highlighted how prompt engineering and few-shot learning improve LLM adaptability to new domains with minimal data [15] and explored MCQ generation for rational pharmacotherapy examinations [16]. A comprehensive review of the AQG methodologies underscores the diversity of approaches, including rule-based, ontology-based, and data-driven techniques, while confirming that ontology-based systems produce semantically rich, context-aware questions [17].

Moreover, integrating structured knowledge representations with natural language generation enables ontology-enhanced LLMs to overcome traditional limitations. One framework has used ontology lexicon induction to improve linked data question answering, resulting in semantically accurate and contextually relevant educational content [18]. Another approach has focused on the development of an automatic question generation system tailored for high school English education, illustrating the effective application of LLMs in structured academic contexts [19].

These examples illustrate the practical advantages of integrating LLMs with ontological structures to enhance both the relevance and instructional alignment of generated content. Broader implications for adaptive and personalized learning are further discussed in studies exploring the evolving role of generative Artificial Intelligence (AI) in education, emphasizing the necessity of prompt engineering, fine-tuning, and semantic grounding [20, 21]. Nevertheless, it is important to note the issue of LLM hallucinations, which hinder the reliability of generated questions, and was critically addressed in [22, 23].

Collectively, these studies emphasize the transformative potential of integrating LLMs with ontology-based knowledge representations. By combining semantic precision with linguistic flexibility, such hybrid systems can produce personalized, adaptive, and domain-specific instructional content. These innovations not only enhance the quality and relevance of generated questions, but also contribute to more scalable, responsive, and effective learning environments. Nevertheless, a significant gap remains in fully understanding the impact of this integration, highlighting the need for novel semantic frameworks that systematically combine LLMs and ontologies. This study addresses the particular gap by proposing an algorithm-driven approach for generating adaptive, domain-aligned educational questions, advancing both automated content creation and targeted instructional support.

The current work explores three methods: template-based structured ontology question generation, LLM-based structured ontology question generation, and LLM-based flat concept list question generation. The contributions of this research are the following :

- Trade-off between structured and unstructured ontologies: The paper identifies a major trade-off in AQG, structured ontologies provide consistency but often lack adaptability to complex or novel settings. A significant contribution is the proposal of hybrid models that combine the generative capabilities of LLMs with the reliability, consistency, and hierarchical linkages of structured ontologies. Future research could enhance the adaptability of structured ontologies by incorporating richer, dynamic textual references and context-specific information, while maintaining the logical coherence necessary for specialized domains (e.g., algorithms, technical subjects). This would expand the capabilities of AQG systems, making them both semantically grounded and adaptable across diverse scenarios.
- Limitations of unstructured ontologies: LLM-based approaches using flat concept lists or unstructured ontologies often produce irrelevant or hallucinated queries, despite achieving strong BERT-based Precision, Recall, and F1-score. This study makes a critical contribution by identifying these limitations. Future work should explore novel strategies to embed contextual grounding into unstructured ontologies, ensuring that generated questions are not only tied to individual concepts, but also contextually coherent. Approaches could include context-aware question generation using broader semantic understanding or leveraging external knowledge bases.

The results highlight the efficiency of different question-generation strategies, particularly in specialized fields, like algorithms, where hierarchical relationships are crucial.

II. METHODOLOGY

A. Framework Design

Three distinct configurations were implemented to evaluate the effectiveness of ontology-based approaches for AQG: i) template-based structured ontology question generation, ii) LLM-based structured ontology question generation, and iii) LLM-based flat concept list question generation.

1) Template-based Structured Ontology Question Generation

The structured ontologies were utilized with predefined templates, for both classes and their corresponding subclasses in the ontology, to automatically generate questions based on hierarchical concepts. These templates were categorized into different types, such as "Definition," "Concept Completion," "Quantification," "Example," and "Feature Specification."

The ontology hierarchies, including relationships, like "is-a" were essential for populating these templates. The questions were generated recursively to explore relationships between concepts and their subclasses. For instance, for the "Concept Completion" category, questions, like "Define {Concept}" or "What is meant by a {Concept}?", were generated for each class and subclass. Similarly, "Example" questions, such as "Can you provide an example of {Concept}?", were created.

Moreover, the question generation algorithm used templates to handle more complex relationships between classes and subclasses. For example, "Concept Comparison" templates were utilized to generate questions comparing different subclasses within a class, such as "How does {Concept1} differ from {Concept2}?" or "What is the difference between

{Concept1} and {Concept2}?". Additionally, "Verification" questions were created to verify the relationship between a class and its subclass, such as "{Concept2} is a type of {Concept1}. True or False?".

An overview of the question types and their corresponding templates is provided in Table I.

TABLE I. QUESTION TYPES AND THEIR CORRESPONDING TEMPLATE QUESTIONS

Question type	Question template	Description logic
Definition	- Define {Concept}. - What is meant by a {Concept}? - Define the following terminology: a){Concept1} b){Concept2}	- ∃ defines.(Concept) - ∃ meaning.(Concept) - {a, b} ⊑ ∃ defines.(Concept)
Concept completion	- Explain what a {Concept} is. - Explain the term {Concept}.	- ∃ explain.(Concept) - ∃ term.(Concept)
Quantification	List the three general {Concept}.	∃hasGeneralConcept.(Concept) ⊓ (count = 3)
Example	- Can you provide an example of {Concept}? - Give an instance where {Concept} can be applied. - Provide an example of {Concept1} and compare it with {Concept2}.	- ∃exampleOf.(Concept) - ∃ instance.(Application ⊓ Concept) -∃ exampleOf.(Concept1 ⊓ Concept2) ⊓ compare(Concept1, Concept2)
Feature specification	- Enumerate the features of {Concept}. - List the characteristics of {Concept}. - Provide two advantages of {Concept}.	- ∃ hasFeature.(Concept) - ∃ hasCharacteristic.(Concept) - ∃ hasAdvantage.(Concept ⊓ (count = 2)) - ∃hasCharacteristic.(Concept)
Interpretation	- How would you interpret {Concept} in the context of {Concept}? - What does {Concept} indicate in {Concept}?	- ∃ interpretInContext.(Concept ⊓ Context) - ∃ indicates.(Concept ⊓ Context)
Verification	- {Concept} is a type of {Concept}. True or False?	Concept1 ⊑ Concept2 ⊔ ¬Concept2
Comparison	- How does {Concept} differ from {Concept}? - What is the difference between {Concept} and {Concept}? - With regard to {Concept}, explain the difference between {Concept} and {Concept}.	- (Concept1 ⊓ ¬Concept2) ⊔ (Concept2 ⊓ ¬Concept1) - ∃ comparisonInContext.(Concept ⊓ (Concept1 ⊓ ¬Concept2) ⊔ (Concept2 ⊓ ¬Concept1))

2) Large Language Model-Based Structured Ontology Question Generation

Structured ontologies were used with LLMs to generate diverse and meaningful questions. The hierarchical structure of the ontology and the relationships between concepts were explicitly encoded into the prompts to guide the LLM's question generation. For example, prompts, such as "You are a helpful question generator. Generate five questions from the concepts mentioned: {context}", were used to leverage the relationships defined within the ontology, enabling the LLM to produce contextually relevant and semantically grounded questions. The LLM produced a variety of question types, including feature specification, true/false, comparison, and interpretation, each informed by the underlying relationships between concepts. For instance, a question could focus on the differences between two related concepts, or it could ask for specific examples or definitions based on the relationships defined in the ontology.

3) Large Language Model-Based Flat Concept List Question Generation

Instead of utilizing a structured, hierarchical ontology, the LLM was provided with a simple list of concepts extracted from the ontology, without predefined relationships. The LLM was tasked with generating questions based solely on these individual concepts, relying on its natural language understanding to interpret the information and produce relevant questions. The process began by prompting the LLM to generate questions for each concept by providing a simple context, such as "Generate 2 questions for the concept [Concept].". These questions were then categorized under different question types, such as "Feature Specification," "True or False," and "Definition."

B. Data Preparation

The data preparation process involved creating both structured and unstructured ontologies, with a particular focus on the domains of algorithms and data structures.

1) Algorithms Ontology

The structured ontology, referred to as the Algorithms Ontology, was created by domain experts [24] and specifically targets the foundational areas of algorithms and data structures. It captures core concepts, relationships, and examples within these domains, forming a comprehensive knowledge base.

The ontology comprises 114 classes and 114 subclass relationships, offering a clear hierarchical organization of algorithmic concepts. Additionally, it defines five object properties that establish relationships between classes and includes 19 concrete instances that illustrate examples of these concepts. Each class represents a specific concept, and subclasses represent more specific instances or variations. As the Algorithms Ontology is bilingual, containing both English and Arabic terms, preprocessing was necessary to ensure consistency. The primary preprocessing step involved extracting only English characters from the ontology, eliminating any Arabic script to streamline the data and maintain uniformity for question generation tasks.

To further enhance the ontology, an augmentation process was implemented using an external knowledge source (GPT-4 mini). This involved adding 138 textual descriptions as annotations to the ontology's classes, properties, and individuals. These detailed textual explanations served as reference material for evaluating the quality and relevance of the generated questions.

2) Flat Concept List

In parallel, an unstructured ontology was also developed. This version was derived from the structured ontology by extracting its concepts and storing them in a simple list format.

Unlike the structured ontology, the unstructured version lacks hierarchical relationships or predefined connections between concepts, and instead, the concepts are presented in a raw, isolated form, and are processed independently for subsequent question generation tasks.

C. Experimental Setup

To evaluate the impact of ontological structure on question generation, the present work designed three distinct experiments comparing structured and unstructured representations of domain knowledge in the context of algorithms. The first approach, template-based structured ontology question generation, utilizes the ontology in its full form, preserving hierarchical relationships, including classes, subclasses, and object properties. The second approach, LLM-based structured ontology question generation, leverages the ontology's hierarchical structure but employs an LLM to generate questions rather than relying on templates. The third approach, LLM-based flat concept list question generation, simplifies the ontology into a flat list of concepts, discarding all structural and relational information. This approach mirrors the common practice of generating questions for individual concepts without considering their semantic context.

By comparing these methods, this study aims to assess the influence of ontological structure on the relevance, diversity, and semantic alignment of the generated questions. The evaluation process involved comparing the generated questions against a set of reference textual descriptions derived from the concepts in the augmented ontology. These descriptions provided detailed information about the concepts but did not include explicit questions. The goal was to measure whether the generated questions accurately reflected the context and meaning conveyed by these descriptions.

D. Evaluation Metrics

To quantitatively evaluate the quality and contextual relevance of the generated questions, two core evaluation approaches were employed. The first involved BERT-based Precision, Recall, and F1-score, which effectively measure lexical alignment and token-level accuracy between the generated questions and the reference descriptions [25]. The second approach utilized Semantic Similarity metrics to assess the conceptual and contextual alignment between the generated questions and the reference descriptions [26]. Each metric was selected for its ability to capture different dimensions of relevance and coherence important for educational question generation.

The BERT Precision is calculated using:

$$\text{Precision}_{\text{BERT}} = \frac{1}{|C|} \sum_{c \in C} \max_{r \in R} \cos(E_c,\, E_r) \quad (1)$$

where C is the set of tokens in the generated question, R is the set of tokens in the reference description, E_c and E_r are their respective contextual embeddings, and $\cos(E_c,\, E_r)$ measures their cosine similarity:

$$\cos(E_c, E_r) = \frac{E_c \cdot E_r}{||E_c|| \times ||E_r||} \quad (2)$$

The BERT Recall is calculated using:

$$\text{Recall}_{\text{BERT}} = \frac{1}{|R|} \sum_{r \in R} \max_{c \in C} \cos(E_r, E_c) \quad (3)$$

The BERT F1-score is calculated using:

$$\text{F1}_{\text{BERT}} = 2 \cdot \frac{\text{Precision}_{\text{BERT}} \cdot \text{Recall}_{\text{BERT}}}{\text{Precision}_{\text{BERT}} + \text{Recall}_{\text{BERT}}} \quad (4)$$

Semantic Similarity assesses the overall alignment in meaning between the generated question and the reference description:

$$\text{Semantic Similarity} = \cos(E_1, E_2) = \frac{E_1 \cdot E_2}{||E_1|| \times ||E_2||} \quad (5)$$

where E_1 and E_2 are the embeddings of the generated question and the reference description, respectively.

By combining these metrics, the evaluation process assessed the semantic alignment of the generated questions with the reference descriptions, ensuring that the questions were contextually meaningful and aligned with the concepts in the Algorithms Ontology.

E. Implementation

For this study, the following tools and technologies were utilized:

- LLM: Mistral-7B-Instruct-v0.1, an instruction-tuned language model [27].
- Libraries: Python programming language, utilizing packages, such as ctransformers for model interaction and BERTScore [28] for evaluation.
- Ontology Tools: Protégé for ontology development and management.

The system setup included:

- Environment: Google Colab, utilizing the T4 GPU runtime for model inference and evaluation.
- Hardware: GPU acceleration provided by the Colab T4 environment was leveraged to efficiently process large-scale question generation tasks.

III. RESULTS

The generated questions were evaluated against reference descriptions to determine the extent to which they reflected the ideas and relationships embodied in the descriptions. Since the reference set consisted of descriptive rather than interrogative material, the evaluation focused on assessing whether the questions faithfully captured the main ideas and context of the reference content. To facilitate this comparison, the performance of each model based on the evaluation metrics is presented in Table II.

The template-based structured ontology question generation approach achieved a Precision of 0.833, a Recall of 0.844, and an F1-score of 0.838, indicating a good balance between relevance and coverage of the generated questions. The Semantic Similarity of 0.563 suggests moderate alignment in meaning with the reference questions. The findings reveal important trade-offs between structured and unstructured representations of domain knowledge. The template-based structured ontology approach generated consistent and interpretable outputs but lacked adaptability for novel or complex questions. In contrast, the LLM-based methods significantly improved relevance and coverage, benefiting from the broader contextual understanding of models, like Mistral-7 B. However, while the unstructured ontology with LLMs achieved higher average scores, it also exhibited a greater incidence of irrelevant or hallucinated questions, due to the absence of semantic relationships. This inconsistency underscores the importance of structured ontologies in ensuring logical coherence and relevance, particularly in domains, like algorithms, where understanding hierarchical relationships is critical. A notable insight is that structured ontologies, when coupled with richer references encompassing multiple concepts, could further improve metrics. like Semantic Similarity, and mitigate the observed trade-offs. Incorporating such references could enable structured ontologies to match the adaptability and relevance of unstructured configurations while preserving their contextual integrity.

TABLE II. SUMMARY OF THE PERFORMANCE EVALUATION

Approach	Evaluation Metrics (Mean score)			
	BERT Precision	BERT Recall	BERT F1-score	Semantic Similarity
Template-Based Structured Ontology Question Generation	0.833	0.844	0.838	0.563
LLM-Based Structured Ontology Question Generation	0.856	**0.863**	**0.859**	0.534
LLM-Based Flat Concept List Question Generation	**0.859**	0.859	**0.859**	**0.567**

Table III combines performance metrics for all three evaluated methods as averages. All approaches demonstrated above-average results, with Precision, Recall, and F1-scores clustering around 0.85, and exhibiting limited variability, as indicated by the low standard deviations (ranging from 0.01 to 0.014). Even at the lower end, all metrics remained robust (above 0.83), reinforcing the dependability and consistency of the methodologies.

The strong performance of the template-based structured ontology method confirms its effectiveness in accurately identifying pertinent questions and covering a wide range of conceptual areas. It also validates the balance achieved between relevance and coverage. However, the moderate Semantic Similarity score underscores a key limitation: while the generated questions are relevant, they may not fully reflect the complexity and richness of the reference content. This limitation arises from the fixed nature of the template structures, which constrain the system's ability to adapt dynamically to varied contexts. Overall, while the template-based approach achieves respectable levels of coverage, balance, and reliability, the results indicate that the more adaptable LLM-based methods demonstrate greater long-term potential for diverse and semantically richer question generation.

TABLE III. TEMPLATE-BASED PERFORMANCE EVALUATION

	Precision	Recall	F1-score
mean	0.85	0.86	0.85
range	0.03	0.02	0.02
std	0.01	0.01	0.01
min	0.83	0.84	0.84
25%	0.84	0.85	0.85
50%	0.86	0.86	0.86
75%	0.86	0.86	0.86
max	0.86	0.86	0.86

The correlation matrix in Figure 1 illustrates the relationships among the evaluation metrics.

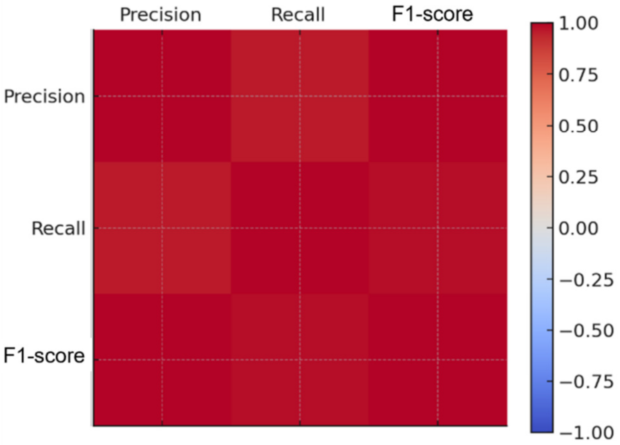

Fig. 1. Correlation matrix.

All metrics exhibit a strong positive correlation, indicating that improvements in Precision are consistently associated with corresponding increases in Recall and F1-score, and vice versa. This strong interrelationship underscores the consistent and balanced performance achieved across the three evaluated techniques.

The results of the confusion matrices for all cases were evaluated. For the template-based structured ontology question generation method, as presented in Figure 2, the number of true positives was 430, indicating that the model successfully identified the mots (86%) of the relevant questions based on structured ontologies and predefined templates. The number of false negatives was 70, suggesting that some relevant questions were missed, likely due to the rigid structure of templates, which may not fully capture nuanced meanings. In addition, the model also produced 50 false positives and 450 true negatives, demonstrating the model's ability to filter out a considerable number of irrelevant questions.

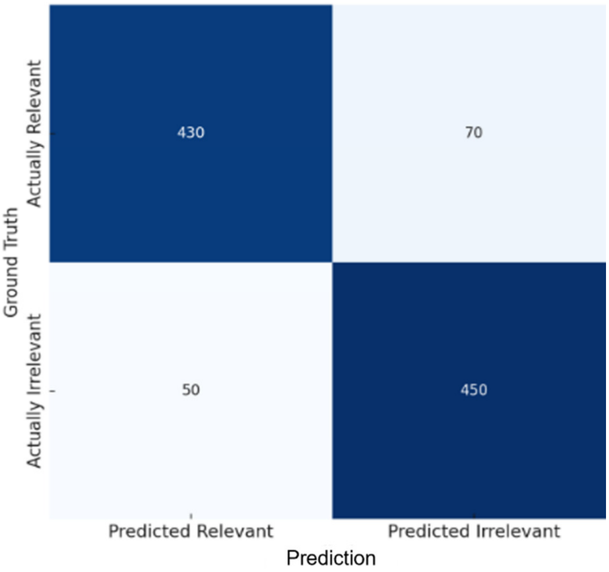

Fig. 2. Confusion matrix.

IV. DISCUSSION

Ontologies provide foundational semantic information, crucial for generating meaningful, practical, and non-trivial questions. Specific elements or related concepts can lead to the creation of highly relevant questions that assess a learner's understanding in depth, while high-level concepts offer broader insights into main issues. Moreover, when concepts are hierarchically organized, inquiries into subclasses can challenge the understanding of the corresponding superclasses. This structure implies that questions targeting closely related concepts are more aligned with the true objectives or needs of the domain being evaluated. Additionally, the knowledge and skills required can be categorized according to different levels of difficulty, reflecting various educational objectives and cognitive demands. Ontologies effectively organize domain-specific knowledge, including concepts, their relationships, and frequently occurring patterns.

The findings reveal that while the template-based structured ontology approach delivers reasonably balanced coverage and performance, LLM-based approaches, owing to their adaptability, demonstrate better general efficacy. This study validates that LLM-based structured ontology question generation is highly effective but highlights the need for improvements in semantic alignment without sacrificing variation. The success of flat concept list generation using LLMs is also confirmed; however, improvements in contextual grounding are proposed to ensure consistent and coherent question generation in structured learning environments.

Performance metrics using Precision, Recall, and F1-score, as displayed in Table II, provide a comprehensive evaluation of the model's effectiveness in question generation. The average scores indicate that the model generates highly accurate (Precision) and comprehensive (Recall) questions. The model's consistency is evident through small standard deviations (approximately 0.01) and narrow ranges (0.02–0.03), demonstrating stability across multiple assessments, as depicted in Table III. The tight alignment of Precision, Recall, and F1-score indicates a strong balance: the model avoids being overly inclusive (producing too many irrelevant questions) or overly conservative (missing relevant ones). The minimal performance variations further reinforce the model's robustness and reliability, showing that it consistently produces accurate, relevant, and broad-coverage questions across different test scenarios.

In Table IV, the performance of the proposed template-based structured ontology question generation is compared against other studies within the same domain.

TABLE IV. COMPARISON WITH RESULTS FROM PREVIOUS STUDIES

Reference	BERT Precision	BERT Recall	BERT F1-score	Semantic Similarity
[5]	0.85	0.86	0.85	0.57
[29]	0.84	0.85	0.84	0.56
[30]	0.83	0.84	0.83	0.55
[31]	0.86	0.87	0.86	0.58
[32]	0.84	0.85	0.84	0.56
This Study	0.83	0.84	0.84	0.56

The performance analysis across studies shows that the study in [29] maintains a balanced performance across all evaluation metrics. The results in [5] demonstrated higher Precision and Recall, leading to an improved F1-score. Among all studies, the work in [30] records the lowest scores across all evaluation metrics, while the research presented in [31] achieved the best overall performance among all compared studies. The model in [32] delivered solid and balanced results across all metrics. Although the performance of the proposed model lies at the lower end compared to other studies, it is important to emphasize that the differences in evaluation metrics between the studies are minimal, not exceeding a 0.03 range.

V. CONCLUSION

In this work, three different methods for Automated Question Generation (AQG) were investigated, utilizing both structured and unstructured ontologies, with and without the integration of Large Language Models (LLMs). The study evaluated template-based structured ontology question generation, LLM-based structured ontology question generation, and LLM-based flat concept list question generation. The employed assessment metrics included BERT Precision, Recall, F1-score, and Semantic Similarity.

The results highlight distinct trade-offs among the approaches. The template-based structured ontology generation method achieved a balanced performance, with a BERT Precision of 0.833, a Recall of 0.844, and an F1-score of 0.838. However, its rigidity limits adaptability to complex or atypical scenarios. The LLM-based structured ontology question generation achieved a BERT Precision of 0.856, a Recall of 0.863, and an F1-score of 0.859. Meanwhile, LLM-based flat concept list question generation, although achieving a BERT Precision, Recall, and F1-score of 0.859, tends to produce hallucinated or less relevant queries due to the lack of semantic structuring.

The findings of this study underscore the critical role of structured ontologies in maintaining logical coherence and semantic relevance, especially in fields, such as algorithms, where understanding hierarchical and relational structures is crucial. This study emphasizes the necessity of a balanced approach, advocating for the enhancement of structured ontologies with richer textual references to improve adaptability without compromising semantic integrity. Furthermore, it proposes that future research explore hybrid models combining the strengths of both structured and unstructured ontologies, applying them across diverse domains to validate and refine the proposed methodologies.

REFERENCES

[1] U. Lee et al., "Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education," *Education and Information Technologies*, vol. 29, no. 9, pp. 11483–11515, Jun. 2024, https://doi.org/10.1007/s10639-023-12249-8.

[2] Z. Wang, J. Valdez, D. Basu Mallick, and R. G. Baraniuk, "Towards Human-Like Educational Question Generation with Large Language Models," in *Artificial Intelligence in Education - 23rd International Conference, AIED 2022, Proceedings*, Durham, United Kingdom, 2022, vol. 13355, pp. 153–166, https://doi.org/10.1007/978-3-031-11644-5_13.

[3] K. Stancin, P. Poscic, and D. Jaksic, "Ontologies in education – state of the art," *Education and Information Technologies*, vol. 25, no. 6, pp. 5301–5320, Nov. 2020, https://doi.org/10.1007/s10639-020-10226-z.

[4] W. Villegas-Ch and J. García-Ortiz, "Enhancing Learning Personalization in Educational Environments through Ontology-Based Knowledge Representation," *Computers*, vol. 12, no. 10, Oct. 2023, Art. no. 199, https://doi.org/10.3390/computers12100199.

[5] D. Nuzzo, E. Vakaj, H. Saadany, E. Grishti, and N. Mihindukulasooriya, "Automated Generation of Competency Questions Using Large Language Models and Knowledge Graphs," in *3rd NLP4KGc @ SEMANTICs 2024*, Amsterdam, Sep. 2024.

[6] K. Nagasaka, "Multiple-choice questions in mathematics: Automatic generation, revisited," in *Proceedings 25th Asian Technology Conference in Mathematics*, vol. 21785, pp. 1-15, 2020.

[7] M. Panahiazar *et al.*, "An Ontology for Cardiothoracic Surgical Education and Clinical Data Analytics," *Studies in Health Technology and Informatics*, vol 294, pp. 407-408, May 2022, https://doi.org/10.3233/SHTI220485.

[8] T. Raboanary and C. M. Keet, "An Architecture for Generating Questions, Answers, and Feedback from Ontologies," in *Metadata and Semantic Research*, vol. 1789, Springer Nature Switzerland, 2023, pp. 135–147.

[9] H. Cheong and A. Butscher, "Physics-based simulation ontology: an ontology to support modelling and reuse of data for physics-based simulation," *Journal of Engineering Design*, vol. 30, no. 10–12, pp. 655–687, Dec. 2019, https://doi.org/10.1080/09544828.2019.1644301.

[10] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 121–204, Mar. 2020, https://doi.org/10.1007/s40593-019-00186-y.

[11] K. Li and Y. Zhang, "Planning First, Question Second: An LLM-Guided Method for Controllable Question Generation," in *Findings of the Association for Computational Linguistics ACL 2024*, Bangkok, Thailand and virtual meeting, 2024, pp. 4715–4729, https://doi.org/10.18653/v1/2024.findings-acl.280.

[12] Y. S. Kıyak, Ö. Coşkun, I. İ. Budakoğlu, and C. Uluoğlu, "ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam," *European Journal of Clinical Pharmacology*, vol. 80, no. 5, pp. 729–735, May 2024, https://doi.org/10.1007/s00228-024-03649-x.

[13] T. Raboanary, S. Wang, and C. M. Keet, "Generating Answerable Questions from Ontologies for Educational Exercises," in *Metadata and Semantic Research*, vol. 1537, Springer International Publishing, 2022, pp. 28–40.

[14] Q. Wang, R. Rose, N. Orita, and A. Sugawara, "Automated Generation of Multiple-Choice Cloze Questions for Assessing English Vocabulary Using GPT-turbo 3.5." *arXiv*, Mar. 2024, https://doi.org/10.48550/arXiv.2403.02078.

[15] G. Agrawal, K. Pal, Y. Deng, H. Liu, and Y.-C. Chen, "CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23164–23172, Mar. 2024, https://doi.org/10.1609/aaai.v38i21.30362.

[16] G. Perković, A. Drobnjak, and I. Botički, "Hallucinations in LLMs: Understanding and Addressing Challenges," in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, Opatija, Croatia, May 2024, pp. 2084–2088, https://doi.org/10.1109/MIPRO60963.2024.10569238.

[17] T. Alsubait, B. Parsia, and U. Sattler, "Generating Multiple Choice Questions From Ontologies: How Far Can We Go?," in *Knowledge Engineering and Knowledge Management*, vol. 8982, Springer International Publishing, 2015, pp. 66–79.

[18] M. DeBellis et al., "Integrating Ontologies and Large Language Models to Implement Retrieval Augmented Generation (RAG)," Applied Ontology, vol. 19, no. 4, pp. 389–407, Jan. 2025.

[19] T. Wang, T. Takagi, M. Takagi, and A. Tamura, "An Automatic Question Generation System for High School English Education," in *Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing*, Mar. 2024, pp. 1215–1219.

[20] M. Al-Yahya, "Ontology-Based Multiple Choice Question Generation," *The Scientific World Journal*, vol. 2014, pp. 1–9, 2014, https://doi.org/10.1155/2014/274949.

[21] S. Maity and A. Deroy, "The Future of Learning in the Age of Generative AI: Automated Question Generation and Assessment with Large Language Models." *arXiv*, Oct. 12, 2024, https://doi.org/10.48550/arXiv.2410.09576.

[22] H. T. Mai, C. X. Chu, and H. Paulheim, "Do LLMs Really Adapt to Domains? An Ontology Learning Perspective," in The Semantic Web – ISWC 2024, vol. 15231, Springer Nature Switzerland, 2025, pp. 126–143.

[23] M. F. Elahi, B. Ell, and P. Cimiano, "LexExMachinaQA: A framework for the automatic induction of ontology lexica for Question Answering over Linked Data," in *Proceedings of the 4th Conference on Language, Data and Knowledge*, Vienna, Austria, Sep. 2023, pp. 207–218.

[24] A. Alnahdi, R. Aboalela, and A. Babour, "Building Bilingual Algorithm English-Arabic Ontology for Cognitive Applications," in *2017 IEEE International Conference on Cognitive Computing (ICCC)*, Honolulu, HI, USA, Jun. 2017, pp. 124–127, https://doi.org/10.1109/IEEE.ICCC.2017.22.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT." *arXiv*, Feb. 24, 2020, https://doi.org/10.48550/arXiv.1904.09675.

[26] J. Wieting, T. Berg-Kirkpatrick, K. Gimpel, and G. Neubig, "Beyond BLEU:Training Neural Machine Translation with Semantic Similarity," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 4344–4355, https://doi.org/10.18653/v1/P19-1427.

[27] *Mistral-7B-Instruct-v0.1-GGUF*. (2024), Mistral AI. [Online]. Available: https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.1-GGUF.

[28] A. H. Nassar and A. M. Elbisy, "A Machine Learning Approach to Predict Time Delays in Marine Construction Projects," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16125–16134, Oct. 2024, https://doi.org/10.48084/etasr.8173.

[29] O. Perera and J. Liu, "Exploring Large Language Models for Ontology Learning," *Issues In Information Systems*, vol. 25, no. 4, 2024, https://doi.org/10.48009/4_iis_2024_124.

[30] A. Lo, A. Q. Jiang, W. Li, and M. Jamnik, "End-to-End Ontology Learning with Large Language Models." *arXiv*, 2024, https://doi.org/10.48550/ARXIV.2410.23584.

[31] Loïc and K. Dassi, "Semantic-Based Self-Critical Training For Question Generation." *arXiv*, Oct. 13, 2021, https://doi.org/10.48550/arXiv.2108.12026.

[32] Z. Wang, K. Funakoshi, and M. Okumura, "Automatic Answerability Evaluation for Question Generation." *arXiv*, Feb. 26, 2024, https://doi.org/10.48550/arXiv.2309.12546.