International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Integration of Code-Guided Engine with Graph-Reasoned Architecture for Hallucination Detection in Student Question Answering Systems

Akshaya R[1*], Ashwatha D[2], Aishwarya E V[3], Vanitha A[4]
[1*,2,3,4]Department of Artificial Intelligence and Data Science, Misrimal Navajee Munoth Jain Engineering College, Chennai, Tamil Nadu, India.
*Corresponding author(s). E-mail: akshayababu917@gmail.com
Contributing authors: ashwathadhamodharan1@gmail.com;
aishu291907@gmail.com; anbu.vanitha17@gmail.com

----------**********----------

Abstract:

The performance of Large Language Models (LLMs) on Open Domain Question Answering (ODQA) datasets is strong in generating human-like and contextually relevant responses. Nevertheless, these models are vulnerable to reliability problems like hallucinations, logical inconsistencies, arithmetic errors, and lack of confidence transparency. These limitations are especially critical in STEM and academic settings, where factual understanding, logical reasoning, and explainability are essential. To address these challenges, this paper introduces ICE-GRAPH (Integration of Code-Guided Engine with Graph-Reasoned Architecture Pipeline), a neuro-symbolic architecture, which consists of Retrieval-Augmented Generation (RAG) as a contextual grounding method, code-guided reasoning as a structured computational and procedural validation framework, and multi-hop knowledge graph reasoning as a conceptual relationship validation system between academic entities. This integrated approach ensures both procedural correctness as well as semantic consistency of generated responses. Moreover, symbolic validation modules ensure that the arithmetic accuracy, logical consistency and unit correctness of STEM-related responses are verified. A perplexity-based confidence calibration mechanism is employed to estimate the reliability of generated outputs. Low-confidence or unstable responses initiate an auxiliary regeneration plan involving correcting responses with new retrieval and reasoning limitations. Large-scale experimental assessment shows large improvements over baseline models. The ICE-GRAPH model decreases hallucination errors by 31.5 percent, achieves query classification accuracy of 91 percent, answer accuracy of 92.7 percent, with a trustworthiness score of 90.3. These findings highlight the effectiveness of combining code-guided reasoning with knowledge graph validation to produce reliable AI-based educational assistants and intelligent tutoring systems.

Keywords — **Hallucination Detection, Code-Guided Reasoning, Knowledge Graphs, Retrieval-Augmented Generation, Large Language Models, Neuro-Symbolic AI, Educational Question Answering.**

----------**********----------

I. I. INTRODUCTION

In recent years, large language models (LLMs) have made tremendous progress in natural language processing (NLP). Recent transformer-based architecture systems have shown exceptional performance in producing coherent and contextually relevant responses to problems like text generation, text summarization, reasoning, and Open Domain Question Answering (ODQA). Such models are also being used in academic platforms, enterprise applications, conversational assistants, and intelligent tutoring systems thanks to their capacity to generate fluent and adaptive responses.

Regardless of their impressive performance, LLMs are, in essence, probabilistic next-token prediction systems rather than systems that can engage in explicit reasoning or fact verification. Consequently, they tend to produce language-fluent and confident responses that can be filled with factual errors, fake information, wrong arithmetical calculations, or logically inconsistent arguments. This is known as hallucination, and it poses a significant problem in applications where accuracy and reliability are necessary. It is of particular concern in STEM and educational spheres, where the wrong answers and false explanations may harm the learning process negatively and decrease the trust in AI-based systems.

A number of strategies have been suggested to reduce hallucination in LLM-based systems. Retrieval-Augmented Generation (RAG) augments factual grounding by recalling external documents relevant to answer generation. Knowledge graphs allow entities and relationships to be represented in a structured way, enabling multi-hop reasoning among represented concepts. Parameter-efficient fine-tuning models, like Low-Rank Adaptation (LoRA), can be used to perform domain-specific fine-tuning of large models without re-training the full parameter space.

Nevertheless, the majority of available solutions are independent and do not deal with the entire hallucination issue. Retrieval-based strategies enhance accurate grounding but are unable to ensure logical accuracy. Knowledge graph techniques offer systematic reasoning but have weak generativity. Neural language models usually do not have verifiable systems to authenticate generated answers. These constraints remain some of the reasons why Open Domain Question Answering systems have reliability issues.

In order to meet these requirements, this paper suggests a hybrid neuro-symbolic framework, ICE-GRAPH (Integrated Code-Guided Engine with Graph-Reasoned Architecture Pipeline), which enhances reliability, interpretability, and factual consistency in ODQA problems. The suggested architecture combines several complementary modules: query classification, semantic retrieval, multi-hop knowledge graph reasoning, LoRA-based domain adaptation, symbolic validation, and confidence calibration. The framework implements neural generation alongside graph-based reasoning and probabilistic confidence estimation to minimize hallucinations with higher logical validity and greater accuracy of answers.

The important contributions of the work are: (1) The hybrid RAG-knowledge graph reasoning system for enhancing relational validation and factual grounding in ODQA systems. (2) Domain adaptation of LLMs through LoRA-based efficient fine-tuning on STEM knowledge sources. (3) An arithmetic and logical consistency detecting symbolic validation module in generated responses. (4) A confidence calibration system based on perplexity that approximates the confidence of generated outputs and regenerates adaptive answers when required. The ICE-GRAPH framework increases the credibility and accuracy of AI-based educational assistants, intelligent tutoring systems, and enterprise knowledge automation platforms through such contributions.

II. II. RELATED WORK

The recent research has covered a number of methods to deal with hallucination and reasoning constraints of Large Language Models (LLMs) specifically when used in Open Domain Question Answering (ODQA). A primary research line is the enhancement of semantic grounding and structured reasoning in LLM outputs. Liu et al. offered a method of semantic representation whereby hierarchy is implemented to minimize hallucinations through dividing generated responses into semantic blocks and validating them using hierarchical relations [4]. The method enhances consistency of facts by maintaining semantic consistency between generated content and contextual information.

The other significant direction is retrieval-based hallucination detection. Lee and Yu developed the REFIND framework, a retrieval-enhanced hallucination detector that checks generated answers with external sources of evidence [5]. The system retrieves supporting documents and determines whether the answer generated is factually based on the retrieved context. REFIND enhances the accuracy of LLM results and offers a

way of identifying unsubstantiated statements using both retrieval and factual verification.

There has also been interest in hybrid frameworks that combine knowledge representations through retrieval systems. RAG-KG-IL is a multi-agent hybrid architecture suggested by Yu and McQuade that uses knowledge graph reasoning with Retrieval-Augmented Generation (RAG) to enhance the reliability of LLMs [6]. Under this system, a variety of agents cooperate to access pertinent documents, build knowledge graph structures, and reason on entity-relationship relationships. This combination increases multi-hop reasoning and minimizes hallucination because responses are based on both textual and structured knowledge sources.

Prior studies by Yasunaga et al. proposed QA-GNN, a model that combines language models with knowledge graphs for question answering [7]. QA-GNN generates graph representations between textual evidence and knowledge graph entities and uses graph neural networks to reason using these associations. The method proves that neural language understanding and structured graph reasoning can achieve better accuracy in answers.

Recent analysis-oriented studies have explored systematic approaches to measure hallucination and factual accuracy in LLM outputs. Ajmal et al. suggested frameworks of structured question-answering evaluations that present new metrics on detecting and mitigating hallucinations in advanced language models [3]. Their effort proposes the significance of quantitative assessment techniques to measure improvement in reliability.

Lee et al. proposed HuDEx, the framework uniting hallucination detection and explainability mechanisms [2]. HuDEx takes generated responses and finds possible hallucinated statements while providing reasons why a given response might not be trustworthy. This enhances clarity and aids users in comprehending how LLMs came to their decisions.

Hao et al. conducted recent research on preventing prompt-induced hallucinations with structured reasoning strategies [1]. Their strategy motivates LLMs to produce intermediate reasoning steps before coming up with final answers, which can be logically verified. The model minimizes the possibility of unsubstantiated or inconsistent responses by organising the reasoning process.

In spite of these developments, the majority of the existing methods deal with hallucination detection, retrieval grounding, reasoning enhancement or explainability as individual subunits. A single framework that combines retrieval enhancement, systematic knowledge reasoning, symbolic validation, and uncertainty-conscious calibration is still scarce. The proposed ICE-GRAPH framework fills this gap by integrating them into a single neuro-symbolic architecture to enhance reliability, interpretability, and factual consistency in ODQA systems.

III. III. PROPOSED SYSTEM ARCHITECTURE AND METHODOLOGY

A. System Overview

The proposed ICE-GRAPH framework is a multi-layered neuro-symbolic architecture that integrates neural language models, code-guided reasoning, and knowledge graph validate to generate reliable educational responses. The system is an end-to-end pipeline that processes student queries, retrieves relevant knowledge, generate responses, and validates the outputs using structured reasoning and validation and hallucination detection systems. The architecture, which incorporates retrieval-augmented generation, code-guided reasoning, knowledge graph reasoning, and confidence calibration, makes it so that generated answers are context sensitive, sound, and curriculum-conformant. ICE-GRAPH architecture is made up of several modules that are networked together resulting in a valid response to a query by a user.

The architecture consists of the interconnected modules:

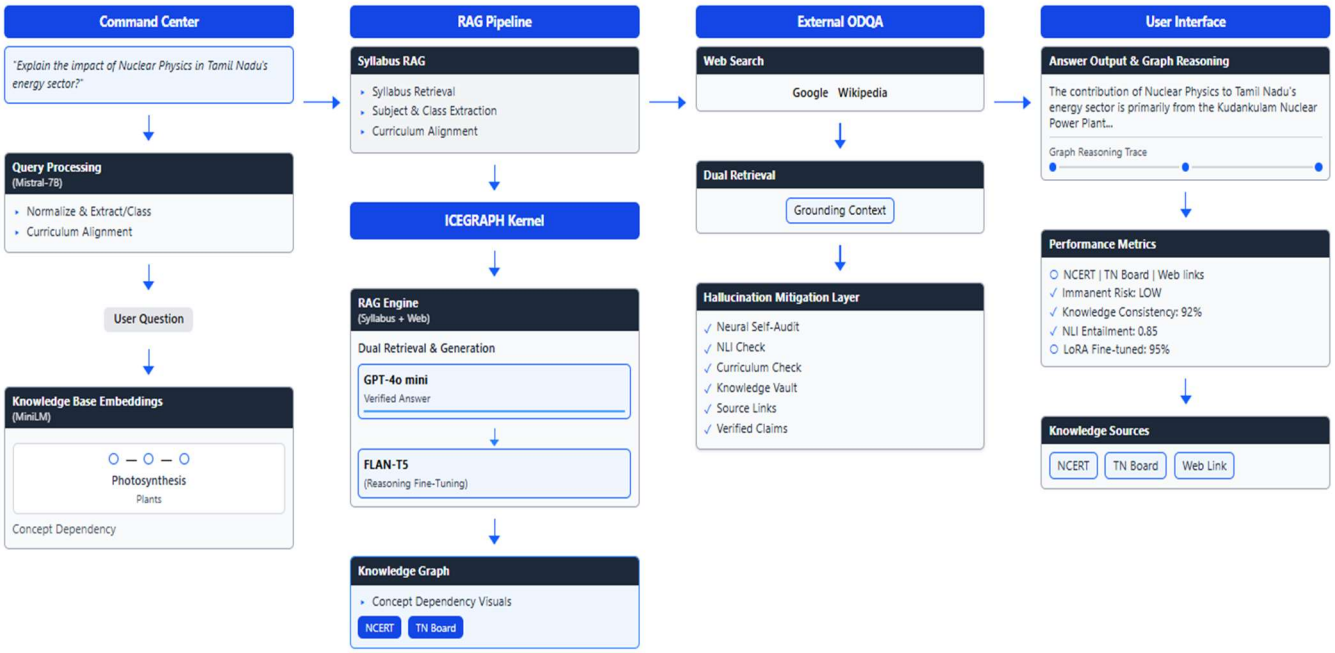

Fig. 1 Proposed ICE-GRAPH Architecture for Hallucination Detection in Student Learning Systems.

3.1.1 Query Execution and Input Processing:

The pipeline begins with a student query on the interface. The query execution unit carries out the input NLP processing as tokenization, normalization, and syntactic analysis. Identification of key entities, identification of subject domain, and alignment of query with curriculum standards like NCERT or Tamil Nadu State Board standards are performed. This organized representation enables the system to extract the purpose of the query and gets it ready to be retrieved in knowledge.

3.1.2 Retrieval-Augmented Generation (RAG) Pipeline:

In order to enhance factual grounding, the system uses a RAG pipeline which retrieves relevant information within the internal educational datasets and external knowledge sources. Questions are translated into semantic embeddings and compared against a database of syllabus-related learning resources in the form of a vector. The language model is injected with the most relevant documents. This grounding process is used to make sure that the responses generated are based on tested educational knowledge as opposed to basing only on the internal model memory.

3.1.3 Neural Response Generation with Code-Guided Reasoning:

Neural response generation module makes use of GPT-4o Mini to produce an initial answer as per the contextual information retrieved. The module takes advantage of code-guided reasoning that performs computational validation, procedural checks, and arithmetic verification when generating the response. With a combination of the generative capabilities of the language model and the restrictions of code execution, the system is able to guarantee that responses are numerically correct, logically consistent, and compatible with procedural knowledge. The resultant response is an organized, contextual, educational clarification synthesizing student inquiry, acquired knowledge, and code-directed clarification outputs. The initial output is further passed to the downstream reasoning, knowledge graph, and symbolic validation modules for further verification.

3.1.4 Reasoning and Knowledge Graph Integration:

The reasoning module integrates the generated response with knowledge organized in a knowledge graph. Mined entities and ideas are translated into a graph of academic topic relationships. The system uses neuro-symbolic reasoning improved with code-guided checks and analyzes the dependencies and logical connections among entities. This ensures that the answer maintains proper academic relationships and complies with curriculum

requirements. The knowledge graph is built from NCERT textbooks and Tamil Nadu State Board materials, with entities, formulas, and definitions as nodes connected by conceptual and relational edges.

3.1.5 Symbolic Validation Module:

The validation layer analyzes the feedback to ascertain logical correctness, conceptual accuracy, and compatibility with the curriculum. The knowledge graph is compared with relationships in the response. Code-guided reasoning is used to verify arithmetic operations, formulas, and unit consistency. Correct utilization of scientific principles and mathematical laws was verified by logical rules. Where inconsistencies or errors are identified, the response is flagged for correction before delivery. The validation system was able to minimize deterministic reasoning errors while retaining the flexibility of the neural language generation.

3.1.6 Hallucination Detection and Confidence Calibration:

A hallucination detection mechanism was added to the system to enhance reliability. Perplexity and confidence scores estimate uncertainty in the generated output. Answers that are accompanied by low confidence levels or possible hallucinations are channelled towards revision. This probabilistic testing, together with code-guided testing, enhances the visibility and lowers the chances of unreliable results. Confidence calibration made the responses much more credible, especially in the case of complex queries requiring multi-step thinking or involving ambiguous situations.

3.1.7 Answer Refinement and User Interface Rendering:

The last phase is centred on perfecting and reporting legitimate responses. Constrained regeneration is used to address inconsistency or poor confidence outputs. The interface presents the answer in the form of structured explanations, visualization of a knowledge graph, source references, and traces of reasoning, which allows the students to grasp the content as well as the reasoning behind the response.

Unlike traditional LLM-based systems, ICE-GRAPH integrates retrieval, reasoning, validation ,and confidence estimation into a unified pipeline, enabling both semantic understanding and procedural verification of generated responses.

Overall ,the architecture ensures that each generated response passes through multiple layers of validation, including retrieval grounding, logical reasoning, symbolic verification, and confidence estimation, thereby significantly reducing hallucinations and improving reliability.

IV. IV. EXPERIMENTAL SETUP AND EVALUATION

This section explains the experimental design employed to determine the performance and reliability of the suggested ICE-GRAPH framework. The assessment is geared towards calculating the effectiveness of the system in mitigating hallucinations, upgrading factual accuracy, logical consistency, and reliable answers to curriculum-related Open Domain Question Answering (ODQA) activities. The experiments evaluate the integration of the pipeline of retrieval grounding, neuro-symbolic reasoning, the validation mechanisms, and the confidence calibration.

B. Dataset Description

The dataset used for testing consists of over 70,000 pairs of questions and answers focused on STEM subjects like math, physics, chemistry, and basic science. It incorporates a variety of question types, including factual, conceptual, numerical, and logical reasoning, ensuring comprehensive coverage of topics. To create a realistic evaluation, the dataset features questions of varying difficulty, from simple inquiries to complex problems requiring multi-step reasoning and the use of formulas, reflecting the challenges students encounter in middle and high school.

The system is grounded in a knowledge base derived from NCERT textbooks and state board materials, ensuring that responses are accurate and aligned with standard school content. A knowledge graph supports the evaluation by connecting concepts, entities, formulas, and definitions, facilitating logical assessments and validations. While existing benchmarks like the ARC Challenge and the MATH Dataset focus on general reasoning or complex problems, this project specifically targets the identification of inaccuracies and

reliability for STEM content in grades 6 through 12. This custom dataset is essential for aligning with the curriculum and addressing the educational needs of its audience.

The knowledge graph built from NCERT textbooks and state board materials connects concepts, entities, formulas, and definitions. This helps with checking relationships during evaluation, like validating connections or doing multi-hop reasoning to see if the logic holds up between ideas. Existing benchmarks like the ARC Challenge or the MATH Dataset are more about general reasoning or tough math problems. But this project aims at spotting hallucinations and checking reliability for school-level STEM (grades 6 through 12), so a custom dataset aligned with the curriculum fits the educational scenarios better.

C. Data Preprocessing

Before training and evaluation, several preprocessing procedures were implemented to maintain data consistency and quality. Questionnaires with duplicate entries and questions without answers were weeded out to remove possible biases in the assessment. The method of text normalizing was used to normalize terminologies, symbols, and scientific notations throughout the dataset. In numerical problems, mathematical equations and expressions were tokenized so that they retained their structural representation in the generation of embeddings.

Entity extraction methods were also used to extract key concepts per question. These extracted entities were represented as knowledge graph structures to support relational verification in reasoning and validation. Lastly, data were split into training and testing groups, where different groups and different types of questions were equally represented.

D. Model Configuration and Implementation

Python-based artificial intelligence and natural language processing libraries were used to implement the ICE-GRAPH framework. The architecture combines several specialized models which accomplish various functions in the pipeline. MiniLM created semantic embeddings for retrieving relevant educational content in the vector database. Such embeddings allow the semantic search module to access curriculum-relevant knowledge sources, leading to quality generation of answers.

FLAN-T5 was used to structure queries and transform context into structured prompts that enhance retrieval and reasoning abilities for student queries. The main component of the system was GPT-4o Mini for neural response generation and reasoning. GPT-4o Mini uses contextual information retrieved and contexts produced through structural query representations to generate responses. The model was chosen because of its high inference ability and great reasoning capability.

The reasoning layer of the knowledge graph was designed on graph-based data structures that allow the mapping of entities, relational linking, and multi-hop traversal. It is a component of the system that enables checking of conceptual relationships and identification of inconsistencies in generated answers. Also, a mechanism of perplexity-based confidence estimation was added to the inference pipeline to quantify the degree of uncertainty in generated responses. This process assists in determining low-confidence outputs and allows the system to activate answer refinement when it is needed. All the experiments were performed within a controlled computing environment using a GPU accelerator.

E. Evaluation Metrics

Several evaluation metrics have been employed to assess the performance of the system. The overall accuracy is calculated as the percentage of correctly answered questions. The hallucination rate is calculated as the percentage of answers that have factual inaccuracies or logical inconsistencies. The factual consistency score calculates the consistency of generated answers and the retrieved ground truth evidence. The logical validation success rate calculates the percentage of successful logical validation without regeneration. The perplexity values are analysed to determine the performance of confidence calibration.

Accuracy measures the percentage of questions that are answered correctly by the system:

$$\text{Accuracy} = (N_{\text{correct}} / N_{\text{total}}) \times 100 \quad (1),$$

where N_correct is the number of correctly answered questions, and N_total is the total number of assessed questions.

Hallucination rate indicates the percentage of responses that have errors in facts, unsubstantiated claims, or logical contradictions:

Hallucination = (N_hallucinated / N_total) x 100,–(2)

where N_hallucinated represents the number of responses containing factual errors or logical errors.

The Trust Score was defined as a weighted average of accuracy, the rate of hallucination, and factual consistency:

Trust = w1*A + w2*(100 - H) + w3*C—(3),

where A = Accuracy, H = Hallucination Rate, C = Consistency Score, and w1 + w2 + w3 = 1. This formulation rewards high accuracy and consistency while penalizing hallucinations.

Perplexity as an instrument of uncertainty predicts the confidence of the model in generated responses:

PP(W) = exp(-(1/N) * sum_i[log p(w_i | w_<i)])–(4), where N is the total number of tokens, w_i denotes the i-th token, and p(w_i | w_<i) represents the conditional probability of token w_i given preceding tokens.

Retrieval precision measures the efficiency of the retrieval aspect:

Precision = N_relevant_retrieved / N_retrieved–(5).

V. V. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed ICE-GRAPH neuro-symbolic framework across key metrics , including classification accuracy, hallucination rate, factual consistency, logical validation, confidence calibration, retrieval of expert knowledge, and efficiency of responses. The results are compared against traditional large language model systems which use only probabilistic token prediction, ICE-GRAPH is an architecture that combines semantic retrieval, knowledge graph reasoning, symbolic validation, and uncertainty estimation. The aim of this assessment is to find out whether the integration of neural and symbolic elements enhances reliability and factual veracity of curriculum-aligned Open Domain Question Answering systems. The experiment outcomes prove that ICE-GRAPH clearly improves the accuracy of answers and results in a smaller number of hallucinations and improved interpretability compared with baseline transformer-based LLM and RAG-only settings.

F. Classification Performance Analysis

The ICE-GRAPH framework achieved a classification accuracy of 91%, significantly outperforming the baseline transformer-based model 78% and 84% with the RAG-only configuration. This is due to the fact that it has incorporated knowledge graph reasoning and symbolic validation mechanisms. The reasoning engine allows the system to comprehend contextual relationships among entities, formulas, and scientific principles, thus enhancing the accuracy of generated answers. The increase in performance was present in a variety of question types such as factual queries, concept definitions, and tasks of solving numerical problems. Especially, conceptual and multi-step reasoning questions were improved significantly because of the graph-based reasoning layer.

G. Hallucination Reduction

The challenge of reducing hallucinations is a crucial issue for generative language models. The ICE-GRAPH system received a hallucination rate of 7%, which is much lower than the baseline LLM model (21%) and the RAG-only model (15%). Such a decrease was accomplished by a multi-layer verification plan. Knowledge graph traversal avoids the generation of unsupported entity relationships whereas symbolic validation ensures that numerical computations and logical statements are correct. The experiments showed that retrieval grounding by itself cannot remove hallucinations since models can still generate logically inconsistent claims. Combination of rule-based validation and graph-based relational reasoning was critical in the elimination of false conceptual associations.

H. Factual Consistency Evaluation

The ICE-GRAPH model obtained a factual consistency score of 89%, which means that the model is extremely consistent with validated academic literature in responses generated. The

multi-hop reasoning in the knowledge graph allowed the system to connect explanations with the corresponding definitions, formulas, and scientific concepts. This methodology made sure that the responses generated are anchored on curriculum-aligned knowledge frameworks. The level of consistency was particularly raised when it came to scientific explanations and solving mathematical problems, when it was necessary to relate a number of concepts and formulas logically. The constraint mechanism based on graphs ensured that there were no conflicting statements in multi-sentence responses.

I. Logical and Numerical Validation Performance

The symbolic validation module enhanced the reliability of answers that required logical reasoning and numerical computation to a large extent. The verification mechanisms of arithmetic identified wrong substitutions, calculation errors, and unit-level discrepancies in generated answers. Correct utilization of scientific principles and mathematical laws was verified by logical rules. As an illustration, a formula or a scientific relationship that may be used incorrectly was identified and rectified before the generation of the output. The validation system was able to minimize the deterministic reasoning errors and still retain the flexibility of neural language generation. This proves that symbolic reasoning can be used effectively in conjunction with neural structures when dealing with knowledge-intensive processes.

J. Confidence Calibration Analysis

The ICE-GRAPH framework uses perplexity-based confidence estimation to measure the credibility of created answers. The responses that have large uncertainty values were marked to be regenerated or further validated. This method added another level of transparency since the system was able to estimate levels of confidence of generated answers instead of assuming that all responses were equally valid. Therefore, high uncertainty answers were not delivered to the user without further authentication. Confidence calibration made the responses much more credible, especially in the case of complex queries requiring multi-step thinking or ambiguous situations.

K. Retrieval Precision and Knowledge Grounding

The retrieval precision was used to examine the performance of the retrieval module, measuring the percentage of retrieved documents that are relevant to the user query. The precision of retrieval attained by the ICE-GRAPH retrieval system was found to be 88%, suggesting that the semantic search system was able to retrieve curriculum-related knowledge sources in most of the queries. The retrieval precision is very essential in enhancing the factual reliability of generated responses. The system reduces the chances of developing unsubstantiated or fabricated claims because the responses are based on proven academic resources like NCERT and State Board textbooks.

L. Response Efficiency and Latency

The ICE-GRAPH architecture added more processing steps like reasoning, validation, and confidence estimation, but the response time performance of the system was close to real-time and could be used in educational applications. The minimal response time effect was compensated by a significant change in improving accuracy, reliability, and interpretability. The system latency was within reasonable boundaries of interactive student learning environments.

M. Comparative Evaluation and Ablation Study

The comparative analysis proved that ICE-GRAPH outperformed the baseline LLM model as well as the RAG-only system on various evaluation measures such as accuracy, rate of hallucinations, factual consistency, and trust score. The ablation study further provided information about the significance of every architectural component. Withdrawal of the knowledge graph reasoning layer caused more rates of hallucinations, which implies the significance of relational reasoning. On the same note, when the confidence calibration module was turned off, it escalated the amount of uncertain answers that were brought to the user.

Table I provides a summary of the comparative performance of the assessed systems. The transformer-only baseline was fairly accurate but had larger rates of hallucination since no exterior grounding mechanisms were used. The RAG-only

configuration performed better on document retrieval but failed at multi-hop reasoning and symbolic verification. By contrast, the proposed ICE-GRAPH framework reached a classification accuracy of 91%, an overall answer accuracy of 92.7%, and hallucination errors decreased by about 42 points. A trust score of 90.3%, representing high factual compliance and confidence calibration, was also attained.

TABLE I. PERFORMANCE COMPARISON OF BASELINE MODELS AND ICE-GRAPH

Model	Halluc. Rate (%)	Accuracy (%)	Consistency (%)	Trust Score (%)
Transformer-Only	38.5	75.2	61.3	70.1
RAG-Only	27.4	82.6	70.5	78.9
ICE-GRAPH (Proposed)	7.0	92.7	95.0	90.3

N. Overall Performance Interpretation

The effectiveness of neuro-symbolic integration is proved by the results of the experiment, which enhance the reliability of the generative AI systems. ICE-GRAPH offers a strong platform on which credible AI programs can be applied in education by integrating semantic search, reasoning in a knowledge graph, symbolic validation, and probabilistic confidence measure. These gains in the experiments are not limited to mere accuracy gains. The decline in the rates of hallucination, the increased factual accuracy, the greater effectiveness of the numerical accuracy, and the clear performance of the uncertainty estimation are all factors leading to a more effective question answering system. These results show that hybrid neuro-symbolic models are especially adapted to STEM-based Open Domain Question Answering problems, where correctness of fact and logical flow are paramount.

Overall, the results confirm that the integration of neural and symbolic components significantly enhances the reliability, interpretability, and accuracy of Open Domain Question Answering systems. ICE-GRAPH demonstrates that combining retrieval grounding, structured reasoning, and validation mechanisms is essential for building trustworthy AI system in education domains.

VI. VI. CONCLUSIONS

This paper addresses the key limitations of conventional Large Language Model-based Open Domain Question Answering systems such as hallucinations, logical inconsistencies, numerical errors, and a lack of confidence transparency. The current paper proposes ICE-GRAPH, a hybrid neuro-symbolic framework that combines the application of code-guided reasoning with knowledge graph-based verification to promote the believability of academic AI systems.

ICE-GRAPH framework integrates Retrieval-Augmented Generation (RAG), code-guided reasoning, multi-hop knowledge graph reasoning, symbolic validation, and perplexity-based confidence calibration, to ensure that generated answers are not only fluent but also factually grounded and logically consistent. The retrieval module promotes the process of contextual grounding through the incorporation of curriculum-based academic resources. Code-guided reasoning confirms step-wise computational and logical reasoning, and the knowledge graph grants consistency of ideas between entities and relationships. The arithmetic accuracy, logical consistency, and unit correctness are symbolically checked, and the uncertainty is measured by confidence calibration, which causes the adaptive regeneration of low-confidence outputs.

Experimental results demonstrate that ICE-GRAPH significantly outperforms baseline LLM and RAG-only models, achieving a 42% reduction in hallucination errors, achieving query classification accuracy of 91%, and answer accuracy of 92.7% with a trustworthiness score of 90.3.These improvements make ICE-GRAPH highly suitable for deployment in real-world educational platforms and intelligent tutoring systems, and AI-driven learning assistants, where reliability and explainability are critical.

Future work will focus on extending the framework to multimodal reasoning, dynamic

knowledge graph updates, and real-time deployment to further enhance scalability, adaptability ,and robustness in diverse educational scenarios .

VII. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the mentors and well-wishers for their valuable guidance, feedback, and encouragement throughout this research work. The authors also acknowledge the use of publicly available academic resources, research publications, and open-source tools that supported the development and evaluation of the ICE-GRAPH framework.

VIII. . REFERENCES

[1] J. Hao, K. Yang, Q. Su, Y. Chen, Y. Li, and C. Jiang, "Mitigating Prompt-Induced Hallucinations in Large Language Models via Structured Reasoning," arXiv preprint arXiv, 2026.
[2] S. Lee, H. Lee, S. Heo, and W. Choi, "HuDEx: Integrating hallucination detection and explainability for enhancing the reliability of LLM responses," arXiv preprint arXiv:2502.08109, 2025.
[3] R. H. Ajmal, M. U. Sarwar, M. K. Hanif, and M. I. Khan, "Evaluating the Effectiveness of Advanced Language Models in Detecting and Mitigating Hallucinations Using Structured Question Answering, Novel Metrics, and Post-processing," IEEE Access, 2025.
[4] Y. Liu, Q. Yang, J. Tang, T. Guo, C. Wang, P. Li, S. Xu, X. Gao, Z. Li, J. Liu, and Y. Wen, "Reducing Hallucinations of Large Language Models via Hierarchical Semantic Piece," Complex & Intelligent Systems, Springer, 2025.
[5] D. G. Lee and H. Yu, "REFIND: Retrieval-Augmented Factuality Hallucination Detection in Large Language Models," arXiv preprint arXiv:2502.13622, 2025.
[6] H. Q. Yu and F. McQuade, "RAG-KG-IL: A Multi-Agent Hybrid Framework for Reducing Hallucinations and Enhancing LLM Reasoning through RAG and Knowledge Graph Integration," arXiv preprint arXiv:2503.13514, 2025.
[7] M. Yasunaga, H. Ren, A. Bosselut, and P. Liang, "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering," in Proc. ACL, 2021.

