AI-Driven Intelligent Tutoring System Using Multi-LLM Orchestration, Retrieval-Augmented Generation, and Knowledge Graphs

1st M.Sai Venkata Durga
Dept. of Artificial Intelligence and Machine Learning
Sasi Institute of Technology and Engineering
Tadepalligudem, India
mangena.durga@sasi.ac.in

2nd Dr.Shaik Mohammad Rafee
Dept. of Artificial Intelligence and Machine Learning
Sasi Institute of Technology and Engineering
Tadepalligudem, India
mohammadrafee@sasi.ac.in

***Abstract*—Intelligent Tutoring Systems (ITS) have received a significant amount of focus due to the development of artificial intelligence in an effort to facilitate learning on a more personalized and interactive level. However, the majority of existing AI-based tutoring systems run on a single large language model and each of them is usually cursory to hallucinating responses, curriculum disconnection, and lack of transparency in reasoning. This paper describes an intelligent tutoring system, which uses AI, with multi-LLM orchestration, including retrieval-augmented generation and knowledge graphs in the context of delivering the correct, definable, and curriculum-aware academic assistance. The proposed system chooses the queries of the learners selectively for special language models and grounds the responses with the assistance of the trusted external resources based on a retrieval-augmented pipeline and uses a curriculum-based knowledge graph to identify the gaps in skills and generate personal learning paths. Additionally, agentic reasoning systems built around ReAct and optimized Tree-of-Thoughts are useful in providing step-by-step reasoning and easy decision-making, and course-specific fine-tuned small language models are useful in increasing domain accuracy and efficiency. The outcomes of the experimental observations made with the assistance of the system-level analysis to determine the effectiveness of personalization, the relevancy of responses and the increased level of engagement with a learner demonstrate the fact that the proposed framework can potentially achieve scalable and reliable intelligent tutoring.**

***Index Terms*—Intelligent Tutoring Systems, Knowledge Graphs, Multi-LLM Orchestration, Agentic Reasoning, Retrieval-Augmented Generation (RAG), Adaptive Learning, Personalized Learning, AI in Education.**

I. Introduction

The booming growth of online learning environments has changed the nature of education delivery in higher education and skill-based educational settings. Online and hybrid learning systems provide flexibility and accessibility, but they do not offer much personalized guidance, continuous feedback, and structured academic support to each individual learner. Due to this, students often lack the ability to follow a sequence of concepts, prerequisite issues, and continuity of learning.

Artificial Intelligence (AI) has been proposed as a potential solution to these issues by the use of intelligent tutoring systems that can modify the learning material to suit the needs of the learners. Recent developments with large language models (LLMs) have made it possible to have conversational tutoring systems capable of answering questions, explaining concepts, and generating learning materials. Nevertheless, most of the existing tutors based on LLM rely on a single generic model and run without a curriculum. This usually causes problems like inaccuracies of facts, illusionary explanations and the lack of consistency with organized academic curriculum.

In order to improve reliability and contextual knowledge, retrieval-augmented generation (RAG) has been suggested as a system to base language model generation on external knowledge sources. Likewise, knowledge graphs have been used within education to reflect the curriculum structure, interrelations between concepts, as well as student advancement. Although these techniques are effective, they are not usually applied in combination with reasoning strategies and learner modeling.

The present paper posits the intelligent tutoring system, which can be developed based on AI, to integrate multi-LLM orchestration, retrieval-augmented generation, and curriculum-based knowledge graphs into a single framework. It is a dynamic system that selects language models that are task-relevant, grounds responses on trusted learning materials and has agentic reasoning, which is based on ReAct and optimized Tree-of-Thoughts, to give transparent and goal-focused tutoring. Moreover, domain accuracy and computational efficiency are improved by course-specific, fine-tuned small language models. When combined, the suggested solution will offer precise, justifiable, and customized academic help that promotes systematic learning and long-term growth of learners.

II. Related Work.

The Intelligent Tutoring Systems (ITS) are not something new and they have embraced the use of artificial intelligence and machine learning. The first works were directed at rule-based tutoring models and adaptive student modeling systems. The article by AlShaikh and Hewahi [1] has provided the critical analysis of AI and machine learning methods in ITS; however, it has concentrated on reinforcement learning, neural

networks, and Bayesian models as an adaptive instructional device. Their study emphasized the importance of modeling learning and learners and feedback processes in order to improve their results of learning.

As a result of the appearance of online education services, dialog-based smart learning assistants are becoming popular. Zobel et al. [2] introduced a Smart Learning Assistant on a MOOC platform that provided individual learning support through the assistance of the conversational AI. Their system had demonstrated increased involvement of the learners but was grounded more on structured dialogue management and not on higher levels of reasoning or the use of multi-modal orchestration. Similarly, Senanayake et al. [3] have reviewed AI assistant systems in programming education and identified the benefits of AI-based help and gamification, as well as the problem of reduced human interaction and overreliance on automation.

Interaction systems based on natural language, such as AutoTutor systems and other dialogue-based systems [4], were effective in promoting student understanding. These systems may be defined as NLP-based dialogue management systems with the ability to provide adaptive explanations and feedback. The conventional conversational ITS models, though, have either a single language model that constrains the extent to which the models are context-grounded and the extent of reasoning.

Most recent developments in AI-based learning assistants are connected with multi-modal communication and personalized recommendations. The concept of a smart learning assistant that integrates speech recognition with large language models was presented by Sangeetha et al. [5] to enable real-time communication and machine summarizing of information. They made access and responsiveness much easier, but they did not employ systematic curriculum modeling and grounding retrieval mechanisms.

Knowledge representation Structured domain modeling has also been experimented with in tutoring systems. Intelligent Teaching Assistant Systems (ITAS): Yacef [6] described that in the ITS architecture, teacher aids, monitoring, and analytics on learners are combined. These systems focus on the importance of systematized elements of pedagogy but do not explicitly integrate large-scale generative models or generative pipelines enhanced by retrieval. These systems are interested in the significance of structured elements of pedagogy yet fail to use generative models at scale or retrieve-enhanced pipelines. Equally, AI-enhanced learning assistant systems [7] are question-generation based, answer-evaluation based, and weak-area-identification based, yet not linked to knowledge graphs to make reasoning that is informed by the curriculum.

To address the problem of hallucination and factual errors in generative systems, a model has been suggested to base model responses on external knowledge sources, and this is RAG. It has been established that the retrieval-based grounding is an important method of increasing the factual truth of knowledge-intensive tasks in NLP (Lewis et al. [8]). The designs of retrieval-based tutoring [9] have been shown to be more covered in the reliability of citation and response compared to generation-only systems in educational settings.

Even more importantly, knowledge graphs have been regarded as structured knowledge to represent curriculum dependencies and the progression of the learner. Research of graph-based educational systems has observed the importance of prerequisite relationships and concept mapping in the development of adaptive learning pathways [10]. However, existing systems tend to utilize knowledge graphs but do not combine them with multi-LLM systems of reasoning.

Despite such developments, conversational AI, knowledge retrieval, curriculum modeling, and adaptive reasoning are still viewed by most of the older systems as independent entities. Literature covering the combined application of multi-LLM orchestration, RAG, knowledge graphs, and agentic reasoning to a single intelligent tutoring system is very sparse. The specified gap causes the proposed methodology, which will include these aspects, to offer credible, explanatory, and curriculum-based instruction.

III. PROPOSED METHODOLOGY

The proposed Intelligent Tutoring System based on AI exists in the form of an integrated, modular system consisting of multi-LLM orchestration, RAG and curriculum-based knowledge graphs and agentic processes of providing trustworthy and personalized academic support. Unlike the conventional intelligent tutoring systems, which use either a single, large language model or operate based on a selected group of canonical rule-based strategies, the proposed methodology combines a range of specialized models and lines of reasoning to enhance the factual accuracy and correspondence to the curriculum and customization. The system will focus on reducing the occurrence of hallucinations, increasing the explainability and providing answers with citations through systematic thinking and justification based on knowledge.

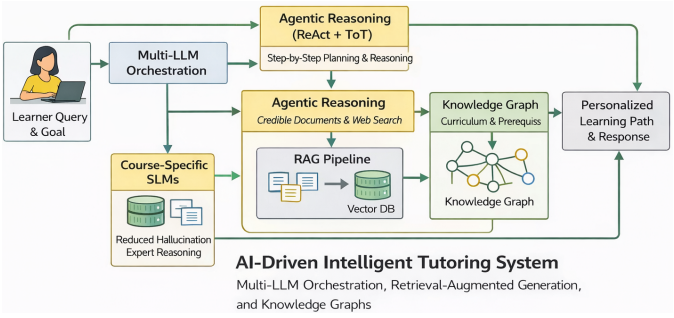

Fig. 1. Overall framework of the proposed AI-driven intelligent tutoring system.

The tutoring process begins by a learner forwarding a query or announcing a learning objective. The multi-LLM orchestration layer receives the query and together decides which reasoning and generation pathway to select based on the difficulty of the task. The agentic reasoning is implemented into the ReAct (Reasoning and Acting) evidence retrieval framework as well as optimized Tree-of-Thoughts (ToT) structured

curriculum planning. A retrieval-enhanced generation pipeline retrieves corresponding contextual information on academic contents that is stored and curated in the form of vectors databases to guarantee responses in terms of facts. Course-specific fine-tuned Small Language Models (SLMs) are also used that facilitate domain-specific explanations and use a curriculum knowledge graph to model topic dependencies and prerequisite relationships to allow the generation of learning paths to be adaptively generated.

Fig. 1 illustrates that the learner query is directed to Multi-LLM Orchestration, the first component that determines whether the task should be conducted in terms of retrieval, reasoning, curriculum planning, or special subject processing. The ReAct-based retrieval and tree of thoughts planning modules work together with the agentic reasoning to break down the intricate problems into the structured reasoning process to facilitate the generation of facts. The RAG pipeline enlists documents and contextual representations of the trusted sources to retrieve the required one. The RAG pipeline will access the documents and contextual embeddings of trusted sources to retrieve the required information to assist in fact-based generation. Simultaneously, the Knowledge Graph likewise possesses inter-module mappings and topic and prerequisite concept mappings, to the extent that in-curriculum progression is known. SLM Course-specific SLMs are correct at the subject level, and the Deep Research Engine verifies answers by grounding them on citations. The resulting system integrates reasoning, retrieval, and curriculum alignment to generate a unique response and a unique course of learning to the learner.

Such a mix of approaches will ensure the scalability, transparency, and reliability of the suggested system can be adjusted to organized academic campuses that require smart tutoring to be adaptive learning and curriculum-related.

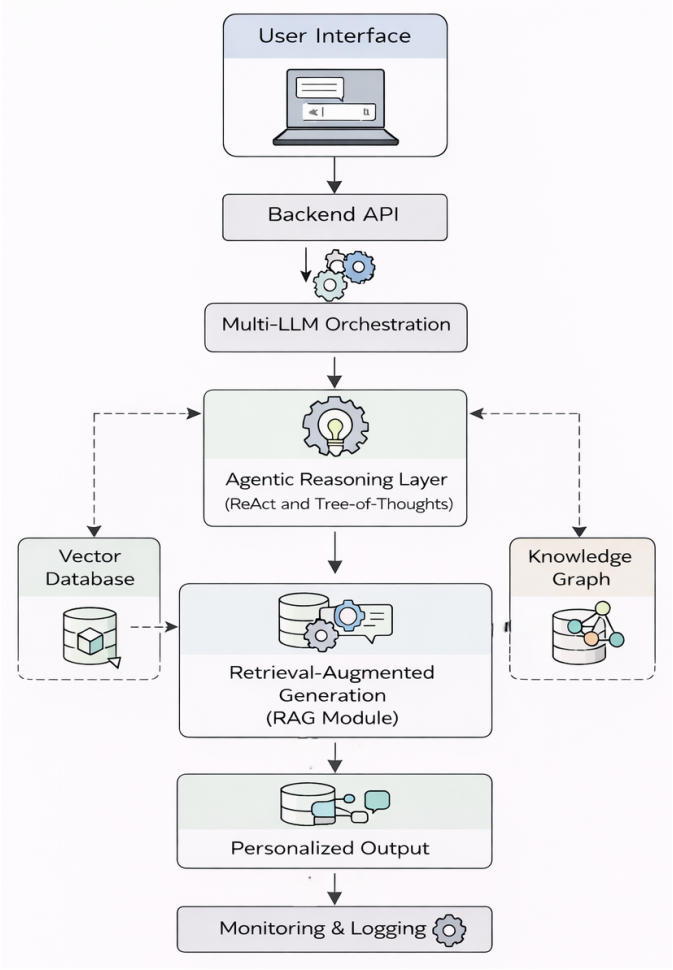

Fig. 2. System Architecture

IV. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The proposed intelligent tutoring system has taken a service-based approach with modular and scalable architecture that provides integration of feature orchestration, reasoning, retrieval, and curriculum modeling into a web-based model. Its architecture is laid out in a stratified style with a user interaction layer, an orchestration and reasoning layer, a knowledge and retrieval layer and a personalization layer.

At the interaction level, learners will be able to access the system through a system of web-based interface, whereby they are able to specify the objectives, make queries, and navigate to the learning system in systematic modes. The front-end communicates through a centralized backend service that receives requests and concentrates the manner in which requests are processed via it, among others. The backend serves as the controlling element since it combines language models and retrieval and graph-based curriculum reasoning modules.

The form of routing controller is used in the multi-LLM orchestration layer, which chooses the appropriate models dynamically as per the task classification. Lightweight, task-specific models are used to do structured reasoning and curriculum planning and more competent, large language models give more sophisticated explanatory answers. The reason why this routing mechanism is applicable is that this routing mechanism is applicable when the selective routing is required, and consequently the computational efficiency remains as high as possible. The agentic reasoning realizes the structured pipelines of prompting that aid the reasoning on ReAct and the planning according to the tree of thoughts. The reason is the controller divides complex academic questions into partial thinking tasks prior to eventually synthesizing reactions.

The RAG module is developed based on a semantic search exploiting a vector database of edited academic texts, i.e., lecture notes and course texts and authenticated external texts. The queries are presented in a form of vectors and compared with the indexed content to identify passages matching the query that are contextually pertinent. The recovered evidence is given to the generation step in such a manner that it will be able to give factual, grounded, and citation-conscious results. The Deep Research Engine also authenticates sources searched via filtering integrity to format trustworthy and

domain-specific sources.

The knowledge graph based on a curriculum is used to model an ordered graph of relationships between modules, topics and prerequisites. The graph database generates the concept dependency, the learning of mastery approximated and the learning of the adaptive path. The gaps in the knowledge demonstrated through the learner interactions that happen as a result of going through the graph make the system propose prerequisite ideas to follow onto higher-level ideas. This will ensure that there is a systematic gradual learning as per the academic curriculum.

It is possible to use 5 domain-aligned fine-tuning course-specific Small Language Models (SLM) to both improve subject-level accuracy and reduce inference latency. These models are applied as domain-specific microservice-style microservices, which could aid in domain-specific explanation, creation of practice questions and concept reinforcement assignments. The reasoning module's retrieval evidence and knowledge graph knowledge outputs are merged in the ultimate response generation pipeline to create an individualized learning suggestion and systematic scholarly response.

The general design is designed to be scalable in a modular manner, to the extent of being able to add new reasoning patterns and new subjects, or even different engines of retrieval, without compromising the basic aspects of the system. Such implementation is scalable, scaled, and suitable to organized educational settings because of the implied approach of its implementation.

V. RESULTS AND DISCUSSION

The proposed AI-driven intelligent tutoring system was evaluated through system-level deployment and interaction analysis across multiple learning sessions. The platform recorded 393 tutor sessions during the evaluation period, with a weekly growth rate of 7% in user engagement. The average number of conversational turns per session was 2.8, indicating structured and focused interactions rather than prolonged unstructured dialogue. Adoption of Tutor Mode showed measurable engagement peaks over a 30-day window, demonstrating consistent learner utilization of guided questioning mechanisms.

The Knowledge Intelligence Profile module tracked 48 total topics per learner, categorizing them into mastered, learning, and struggling concepts. The system dynamically identified weak areas such as recursion-related concepts and automatically generated targeted reinforcement activities. Mastery progression was observed through skill-tree advancement and completion of structured assessments. The Skill Tree Adventure interface enabled sequential unlocking of topics, reinforcing prerequisite-based progression modeled through the curriculum knowledge graph.

Gamification mechanisms contributed to measurable engagement improvements. Users earned experience points (XP), unlocked badges such as "Boss Slayer," and progressed across levels through structured challenges. For example, completion of boss battles with a 60% pass threshold awarded 40 XP and triggered level advancement. The gamification dashboard recorded 187 total XP awarded, 2 badges earned, and active streak tracking, indicating sustained learner motivation. The badge collection and XP-based leveling system provided positive reinforcement loops without compromising academic structure.

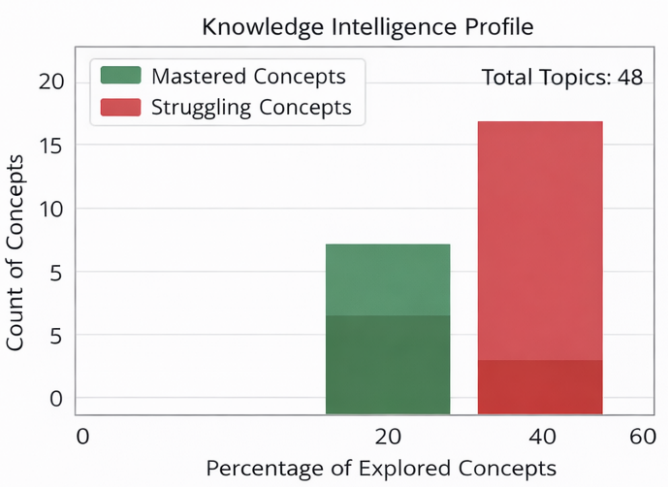

Fig. 3. Knowledge intelligence profile showing topic-level mastery tracking.

The admin dashboard further demonstrated system scalability and monitoring capabilities. Analytics included total tutor sessions, weekly growth trends, average session depth, and content gap analysis. The system supports document upload and identifies unanswered student queries, enabling instructors to refine learning materials. This validates the integration of monitoring and adaptive content feedback within the architecture.

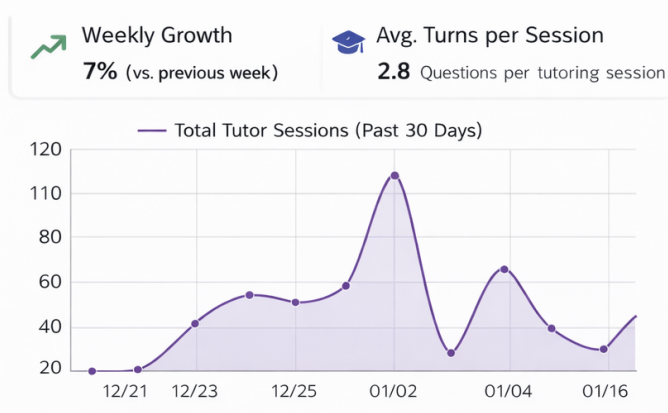

Fig. 4. User engagement metrics and session activity analysis

Overall, the results indicate that integrating multi-LLM orchestration, agentic reasoning, retrieval grounding, and knowledge graph modeling produces a structured and adaptive learning environment. The system effectively identifies learner weaknesses, promotes mastery-based progression, and maintains engagement through gamification. While large-scale

quantitative benchmarking remains future work, the current deployment demonstrates functional reliability, curriculum alignment, and scalable interaction management within an intelligent tutoring framework.

TABLE I
COMPARISON BETWEEN EXISTING APPROACHES AND PROPOSED AI TUTOR ITS FRAMEWORK

Aspect	**Existing Approaches**	**Proposed Framework**
Architecture	Single LLM or rule-based systems	Multi-LLM orchestration with agentic reasoning
Factual Grounding	Limited; prone to hallucination	RAG-based retrieval with citation support
Curriculum Alignment	Generic content responses	Knowledge graph-based structured progression
Personalization	Basic prompt-level adaptation	Mastery tracking with adaptive learning paths
Engagement Support	Static interaction	Gamification (XP, badges, challenges)
Analytics	Minimal monitoring	Admin dashboard with session insights

VI. CONCLUSION

The proposed AI-driven intelligent tutoring system demonstrates how the integration of multi-LLM orchestration, retrieval-augmented generation, structured knowledge graphs, and agentic reasoning can transform traditional conversational AI into a goal-oriented learning companion. Unlike standalone chatbot-based tutors, the system combines curriculum modeling, reasoning transparency, and adaptive personalization within a unified architecture. The deployment results indicate consistent learner engagement, measurable mastery tracking, and structured progression through skill-tree modeling and gamified reinforcement mechanisms.

By incorporating course-specific small language models and a citation-aware deep research engine, the framework enhances factual reliability while maintaining domain alignment. The agentic reasoning layer enables guided Socratic interaction, supporting conceptual clarity rather than surface-level answer delivery. Additionally, the analytics-enabled admin dashboard provides scalable monitoring, session-level insights, and content gap identification, demonstrating the practical viability of the system in academic environments.

Overall, the framework presents a scalable and modular intelligent tutoring architecture capable of balancing reasoning depth, factual grounding, and learner engagement. Future work will focus on large-scale empirical validation, automated knowledge graph expansion, and optimization of orchestration strategies to further enhance efficiency and educational impact.

REFERENCES

[1] F. AlShaikh and N. Hewahi, "AI and Machine Learning Techniques in the Development of Intelligent Tutoring System: A Review," in *Proc. 2021 Int. Conf. on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 2021, doi: 10.1109/3ICT53449.2021.9582029.

[2] T. Zobel, T. Staubitz, and C. Meinel, "Introducing a Smart Learning Assistant on a MOOC Platform: Enhancing Personalized Learning Experiences," in *Proc. IEEE 2nd German Education Conf. (GECon)*, 2023, pp. 1–6, doi: 10.1109/GECON58119.2023.10295099.

[3] S. Senanayake, K. Karunanayaka, and K. V. J. P. Ekanayake, "Review on AI Assistant Systems for Programming Language Learning in Learning Environments," in *Proc. 8th SLAAI Int. Conf. on Artificial Intelligence (SLAAI-ICAI)*, 2024, doi: 10.1109/SLAAI-ICAI63667.2024.10844969.

[4] P. Patel *et al.*, "AI-Powered Chatbots as Virtual Teaching Assistants in Online Education Using Natural Language Processing for Student Support," 2025.

[5] M. M. Sangeetha, N. Nelakurthi, P. K. Anumala Setty, D. Gingupalli, and C. Katta, "AI-Powered Smart Learning Assistant with Speech Recognition," in *Proc. 4th Int. Conf. on Automation, Computing and Renewable Systems (ICACRS)*, 2025, doi: 10.1109/ICACRS67045.2025.11324149.

[6] K. Yacef, "Intelligent Teaching Assistant Systems," in *Proc. Int. Conf. on Computers in Education (ICCE)*, 2002, pp. 136–143.

[7] N. Sai Sumanth, S. Vishnu Priya, S. M. Sankari, and K. S. Kamatchi, "AI-Enhanced Learning Assistant Platform," in *Proc. 7th Int. Conf. on Inventive Computation Technologies (ICICT)*, 2024, doi: 10.1109/ICICT60155.2024.10545011.

[8] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[9] T. H. Chen *et al.*, "Retrieval-Augmented Tutoring for Factual and Citation-Grounded Learning," 2023.

[10] S. Ji *et al.*, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2022.