JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A SELF-VALIDATING AND SELF-REGULATING ARTIFICIAL INTELLIGENCE TEACHING ASSISTANT

[1]Abhiraj Mohan Kadam, [2]Nirbhay Singh
[1]MSc.Computer Science,[1]Student, [2]Professor
[1,2]Nagindas Khandwala College
[1,2]University of Mumbai, Maharashtra, India

***Abstract*:** The rapid integration of Artificial Intelligence (AI) into education has transformed the way learners access information, receive feedback, and engage with instructional content. AI-driven Teaching Assistants (AI-TAs) have emerged as scalable tools capable of answering student queries, supporting personalized learning, automating grading, and facilitating interactive learning experiences. Although recent studies suggest that AI-TAs can match or even surpass human teaching assistants in response speed, availability, and consistency, several critical challenges remain unresolved. These include the risk of hallucinated content, lack of accountability, algorithmic bias, pedagogical misalignment, and the absence of internal quality assurance mechanisms. Most existing AI-TAs are designed to optimize linguistic fluency and task completion rather than epistemic reliability and ethical compliance.

This paper proposes a novel framework for a Self-Validating and Self-Regulating Artificial Intelligence Teaching Assistant (SVSR-AI-TA) that embeds reflective and governance capabilities directly into the system. The framework introduces two core layers: a self-validation layer that verifies factual accuracy, confidence, and consistency, and a self-regulation layer that enforces ethical, and institutional constraints. Over-reliance on AI-generated solutions may reduce learners' critical thinking and problem-solving skills [1]. Drawing upon a synthesis of fifteen peer-reviewed studies in Artificial Intelligence in Education (AIEd), this work identifies structural limitations in current AI-TAs and demonstrates how internal governance mechanisms can address these gaps. A multi-layered system architecture is presented in which generative modules operate alongside validation engines, regulatory controls, and adaptive feedback loops. The paper argues that future educational AI systems must transition from reactive tools into reflective learning partners capable of evaluating and regulating their own behavior. The proposed framework aims to enhance trust, fairness, learner engagement, and institutional scalability in higher education.

***Index Term*:** Artificial Intelligence in Education, AI Teaching Assistant, Self-Validation, Self-Regulation, Generative AI, Retrieval-Augmented Generation, Ethics, Adaptive Learning.

I. INTRODUCTION

The digital transformation of education has accelerated significantly over the past decade, driven by the widespread adoption of online learning platforms, massive open online courses, and blended learning environments. Universities and training institutions worldwide now face unprecedented challenges, including rapidly increasing student enrollments, shortages of qualified instructors, and the growing demand for personalized learning experiences. Two-stage training with active signals yields higher accuracy than one-stage distillation [3]. These pressures have intensified the need for scalable instructional support systems capable of assisting learners without compromising educational quality. Artificial Intelligence Teaching Assistants (AI-TAs) have emerged as promising solutions, offering automated question answering, adaptive tutoring, feedback generation, and learning analytics at scale.

Recent advances in large language models have enabled AI systems to engage in human-like dialogue, reason across topics, and generate coherent explanations. Empirical studies show that learners often complete tasks more efficiently with AI-TAs and report satisfaction levels comparable to interactions with human assistants. However, despite these benefits, AI-TAs remain fundamentally limited. They can generate factually incorrect responses, exhibit biases, violate privacy norms, and fail to align with pedagogical objectives. Proposed artificial intelligence algorithm and deep learning techniques for development of higher education [9]. Most importantly, current AI-TAs lack internal mechanisms to verify their own outputs or regulate their behavior, making them reactive systems rather than accountable educational partners.

This paper addresses these limitations by proposing a Self-Validating and Self-Regulating Artificial Intelligence Teaching Assistant (SVSR-AI-TA). Unlike conventional systems that deliver raw generative outputs directly to learners, the proposed framework introduces a meta-cognitive layer that evaluates and governs responses before they are delivered. By embedding confidence estimation, retrieval-based verification, ethical compliance, and pedagogical alignment within the AI-TA itself, the system becomes capable of self-reflection and self-governance.

The primary contributions of this work are threefold. First, it synthesizes existing AI-TA research to identify design and governance gaps. Second, it introduces a conceptual architecture for self-validation and self-regulation. Third, it outlines design principles for creating ethical, scalable, and learner-centered AI-TAs.

II. RELATED WORK

A. Generative AI as Teaching Assistants

Generative AI systems have demonstrated strong potential in supporting learning environments, particularly in introductory programming and large online courses. Studies indicate that conversational AI can guide novice learners through structured problem-solving strategies, enabling them to complete tasks faster while maintaining high levels of engagement. These systems provide immediate feedback, reduce waiting time for assistance, and scale effortlessly across thousands of learners. However, while these results highlight the efficiency of AI-TAs, they also reveal a dependency on model training quality and data relevance, which may vary across domains and institutions.

B. Confidence and Uncertainty Modelling

One of the most significant limitations of generative AI is its tendency to produce overconfident responses even when uncertain. Research in confidence-aware learning demonstrates that uncertainty estimation can serve as an internal reliability signal. AI significantly improves grading efficiency and feedback consistency in education [10]. By quantifying how confident a system is in its outputs, AI models can identify potentially unreliable responses and trigger corrective actions. This principle forms the foundation of self-validation, as it allows AI-TAs to recognize when they may be wrong rather than blindly responding.

C. Retrieval-Augmented Assistants

Retrieval-Augmented Generation (RAG) has emerged as a powerful technique to ground AI responses in verified knowledge sources. By retrieving relevant documents from curated databases and using them to guide generation, RAG-based systems significantly reduce hallucinations and improve domain accuracy. Custom GPT assistants can reduce technical barriers and democratize AI use in education [5]. Educational studies show that RAG-enhanced AI-TAs outperform purely generative systems in terms of relevance, consistency, and factual correctness, making retrieval a critical component of any reliable educational AI.

D. Personalized and Adaptive Learning

Adaptive learning systems leverage learner data to tailor content, feedback, and pacing. Machine learning models can predict learning styles, identify dropout risks, and recommend personalized pathways. These systems demonstrate that AI can move beyond static tutoring toward dynamic learning partnerships. Cosine similarity-based embedding retrieval improves contextual accuracy in AI-driven information systems [6]. However, personalization without regulation may reinforce biases or foster over-dependence, highlighting the need for governance mechanisms.

E. Ethical and Pedagogical Challenges

Ethical concerns such as bias, privacy, and academic integrity remain central to the debate on educational AI. Scholars emphasize that AI systems must not only be technically efficient but also socially responsible and pedagogically aligned. Without regulatory layers, AI-TAs risk undermining fairness, autonomy, and trust.

III. SELF-VALIDATION AND SELF-REGULATION AI-TA FRAMEWORK

A. Definition and Conceptual Foundations

A Self-Validating and Self-Regulating Artificial Intelligence Teaching Assistant is an advanced educational system designed not only to generate instructional content but also to continuously evaluate and govern its own behaviour before interacting with learners. Strategic integration is required, as AI TAs may provide excessive guidance and code compared to human Tas [1]. Unlike conventional AI-TAs that prioritize fluency, speed, or surface-level relevance, this framework introduces reflective capabilities that allow the system to assess correctness, fairness, ethical safety, and pedagogical suitability prior to response delivery. The AI-TA therefore becomes an autonomous yet accountable learning partner that is capable of self-monitoring and self-correction.

This framework is grounded in two interdependent governance principles.

1. Self-Validation refers to the system's capacity to verify the accuracy, consistency, and reliability of its outputs using internal confidence checks, retrieval verification, and model consensus mechanisms. Stiennon et al. (2020) introduced human feedback–based learning for better text generation [2]. Through this process, the AI system critically evaluates whether its response is factually supported, logically coherent, and appropriate for the learner's context.

2. Self-Regulation refers to the system's ability to enforce ethical, institutional, and pedagogical constraints. This includes detecting bias, protecting learner privacy, preserving academic integrity, and aligning all assistance with instructional goals.

Together, these two principles transform the AI-TA from a reactive information generator into a reflective, accountable, and responsible learning system. Vision, challenges, roles and research issues of artificial intelligence in education [11].

IV. SELF-VALIDATION LAYER

The self-validation layer functions as the internal quality assurance core of the SVSR-AI-TA. Its purpose is to ensure that every system-generated response meets minimum standards of accuracy, reliability, coherence, and contextual relevance before being presented to learners. D. Mpini, Application of artificial intelligence for virtual teaching assistance, Introduction to Information Technology [8]. Traditional AI-TAs rely entirely on their pre-trained knowledge and probabilistic reasoning, which exposes students to the risk of hallucinations, conceptual errors, and misleading explanations. The self-validation layer introduces a meta-cognitive verification stage that enables the system to evaluate its own outputs in real time.

A. Confidence Estimation and Uncertainty Modeling

The validation process begins with confidence estimation, in which the system generates multiple candidate responses using stochastic decoding strategies. Pattern Recognition and Machine Learning [14]. These outputs are compared using semantic similarity and logical consistency measures. When the responses converge on the same explanation, the system infers high confidence in the generated answer. However, when substantial variation is detected, this signals uncertainty or ambiguity. In such cases, the AI-TA does not immediately respond but instead triggers corrective actions, such as requesting clarification from the

learner, retrieving additional knowledge, or flagging the query for review. This process ensures that uncertainty is acknowledged rather than hidden.

B. Knowledge Verification through Retrieval

Once a response passes the confidence threshold, it undergoes knowledge verification using retrieval-augmented generation. The system retrieves relevant material from trusted academic sources, such as course notes, textbooks, and institutional repositories, and compares the generated response against this content using semantic entailment and contradiction detection models. Exploring opportunities and challenges of artificial intelligence in higher education institutions [7]. If discrepancies are identified, the system revises or regenerates the response using verified evidence. This grounding mechanism ensures that explanations are not merely fluent but factually supported and academically reliable.

C. Model Consensus and Cross-Verification

To further strengthen reliability, the system employs a model consensus mechanism in which multiple AI models independently process the same query. Their outputs are compared for factual agreement and logical consistency. When a majority of models converge, the response is approved. An intelligent tutoring system for software engineering courses [15]. When disagreement occurs, the system interprets this as an indicator of ambiguity or insufficient knowledge and reinitiates the validation cycle. This ensemble strategy mirrors academic peer review, minimizing individual model bias and error.

D. Learner Feedback Integration

After each interaction, learners are encouraged to rate the usefulness, clarity, and correctness of responses or provide corrections. LLMs such as GPT, LaMDA, and LLaMA provide more context-aware chatbot responses [4]. These signals are analyzed to refine confidence thresholds, improve retrieval ranking, fine-tune generation behavior, and identify recurring misconceptions. Over time, the system adapts to learner needs, improving both trustworthiness and instructional effectiveness.

V. SELF-REGULATION LAYER

The self-regulation layer governs how the AI-TA behaves, ensuring that all interactions comply with ethical standards, institutional policies, and pedagogical objectives. Effective use of AI requires well-designed prompts and digital competence from educators [5]. While the validation layer ensures that content is correct, the regulation layer ensures that it is delivered responsibly and appropriately.

A. Ethical Controls

The system continuously monitors its language for biased, discriminatory, or harmful content. It also enforces privacy protection mechanisms and data minimization policies to prevent misuse of learner data. Intent classification chatbots help students with academic decision-making [4]. Through fairness auditing and content moderation rules, the AI-TA actively corrects problematic outputs rather than assuming neutrality.

B. Pedagogical Alignment

Rather than simply delivering direct answers, the AI-TA adapts its explanations to match instructional goals. It scaffolds difficult concepts, guides learners through reasoning steps, and adjusts difficulty based on performance. Integration of AI systems with LMS platforms improves adaptive learning and student engagement [6]. This ensures that assistance promotes conceptual understanding rather than passive consumption.

C. Academic Integrity Protection

To preserve academic honesty, the system limits direct solution disclosure for graded tasks. Instead, it provides hints, step-by-step reasoning prompts, and reflective questions. Zhong et al. (2022) developed a multi-dimensional evaluator for text generation quality [2]. This encourages problem-solving and prevents overreliance on automation.

VI. INTEGRATED GOVERNANCE CYCLE

The SVSR-AI-TA operates through a continuous governance loop: Generate → Validate → Regulate → Deliver → Learn → Improve. Each response is generated, verified for correctness, regulated for ethical and pedagogical compliance, delivered to the learner, evaluated through feedback, and used to update system behavior. Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods [12]. This closed-loop design enables continuous improvement while maintaining accountability and trust.

VII. DISCUSSION

The SVSR-AI-TA framework represents a paradigm shift in educational AI design. By embedding internal validation and regulation mechanisms, the system mitigates hallucination, reduces bias, and ensures pedagogical alignment. This transforms AI-TAs from passive tools into self-governing educational agents. TA-Teacher and student internal signals significantly boost data quality and accuracy [3]. The framework supports a transition from AI-directed learning, where students passively receive information, to AI-empowered learning, where learners actively engage in guided knowledge construction.

VIII. LIMITATIONS AND FUTURE WORK

A. Computational Constraints

Multi-stage validation and model consensus introduce computational overhead that may limit real-time deployment. Ethical implications and principles of using artificial intelligence models in the classroom [13]. Future research should explore lightweight validation models and optimization strategies.

B. Interpretability Challenges

Although reliability improves, the internal reasoning processes remain partially opaque. Explainable AI methods are needed to clarify why responses are accepted, modified, or rejected.

C. Future Research Directions

1. Emotional intelligence integration to detect learner frustration and disengagement.
2. Cross-cultural fairness models to ensure linguistic and cultural inclusivity.
3. Human-in-the-loop governance allowing educators to guide system policies.

4. Longitudinal learning analytics to measure long-term educational impact.

IX. CONCLUSION

This paper proposed a Self-Validating and Self-Regulating Artificial Intelligence Teaching Assistant framework that embeds internal mechanisms for verification, ethical governance, and pedagogical alignment. By transforming AI-TAs into reflective, accountable learning partners, the framework enhances trust, fairness, and educational effectiveness, enabling responsible AI adoption in higher education.

X. REFERENCES

[1] Changyoon Lee, Junho Myung, Jieun Han, Jiho Jin, Alice Oh, (2023), "Learning from Teaching Assistants to Program with Subgoals: Exploring the Potential for AI Teaching Assistants" Link
[2] Anmol Agarwal, Yann Hicke, Qianou (Christina) Ma, Paul Denny, (2023), "AI-TA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs" Link
[3] Yuhang Zhou, Wei Ai, (2024), "Teaching-Assistant-in-the-Loop: Improving Knowledge Distillation from Imperfect Teacher Models in Low-Budget Scenarios" Link
[4] Bashaer Alsafari, Eric Atwell, Aisha Walker, Martin Callaghan, (2024), "Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants" Link
[5] Antonio Julio López-Galisteo, Oriol Borrás-Gené, (2025), "The Creation and Evaluation of an AI Assistant (GPT) for Educational Experience Design" Link
[6] Ramteja Sajja, Yusuf Sermet, Muhammed Cikmaz, David Cwiertny, Ibrahim Demir, (2024) "Artificial Intelligence-Enabled Intelligent Assistant for Personalized and Adaptive Learning in Higher Education" Link
[7] Yi Liu, Zerui Yao, (2022), "The application of artificial intelligence assistant to deep learning in teachers' teaching and students' learning processes" Link
[8] Obert Muzurura, Tinomuda Mzikamwi, Taurai George Rebanowako, Dzinaishe Mpini (2023), "Application of artificial intelligence for virtual teaching assistance" Link
[9] Amin Al Ka'bi, (2022), "Proposed artificial intelligence algorithm and deep learning techniques for development of higher education" Link
[10] Soni Maitrik Chandrakant, (2025), "AI-powered teaching assistants: Enhancing educator efficiency with NLP-based automated feedback systems" Link
[11] Fan Ouyang, Pengcheng Jiao, (2021), "Artificial intelligence in education: The three paradigms" Link
[12] Firuz Kamalov, David Santandreu Calonge, Ikhlaas Gurrib, (2023), "New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution" Link
[13] Marc Alier, Francisco José García-Peñalvo, Jorge D. Camba, (2024), "Generative Artificial Intelligence in Education: From Deceptive to Disruptive" Link
[14] Shadeeb Hossain, (2025), "Using Artificial Intelligence to Improve Classroom Learning Experience" Link
[15] Asmar Ali, Andreas Deuter, (2023), "An AI assistant for education in automation" Link