JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8319

OPTIMISING EXPLAINABLE AI IN EDUCATION: A DSPY-BASED FRAMEWORK WITH CHAIN-OF-THOUGHT REASONING FOR ADAPTIVE LEARNING

Ben Rahman[1*]

Information Systems[1]
Universitas Nasional, Jakarta, Indonesia[1]
www.unas.ac.id[1]
benrahman@civitas.unas.ac.id[*]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *The application of artificial intelligence (AI) in education is often constrained by limited reasoning transparency, computational demands, and reproducibility challenges. This study proposes and conducts an exploratory evaluation of a modular AI education framework based on Declarative Structured Programming (DSPy) and Chain-of-Thought reasoning. The framework integrates typed input–output signatures with structured inference to support transparent question generation and adaptive feedback. A pilot study involved 10 participants with computer-science or education backgrounds; each completed three sessions, yielding 30 session-level observations. The framework used GPT-4o mini and was compared with prompt-based and rule-based baselines. It achieved 92.4% session-level learning accuracy and higher observed reasoning-clarity ratings than the baselines, with an average response time of 1.2 s. Because the sample was small, technically oriented, and evaluated over short sessions, the findings constitute preliminary evidence and should not be generalized to diverse learners or sustained learning outcomes. Larger, heterogeneous, longitudinal, and resource-instrumented studies are required.*

Keywords*: Artificial Intelligence in Education, Chain-of-Thought Reasoning, DSPy, Explainable AI, Modular Learning Systems.*

Intisari—*Penerapan kecerdasan buatan (AI) dalam pendidikan sering menghadapi keterbatasan transparansi penalaran, kebutuhan komputasi, dan reproduktibilitas. Penelitian ini mengusulkan serta mengevaluasi secara eksploratif framework AI edukasi modular berbasis Declarative Structured Programming (DSPy) dan Chain-of-Thought. Framework mengintegrasikan input–output bertipe dengan inferensi terstruktur untuk mendukung pembuatan pertanyaan dan umpan balik adaptif yang transparan. Studi pilot melibatkan 10 peserta berlatar belakang ilmu komputer atau pendidikan; setiap peserta mengikuti tiga sesi sehingga diperoleh 30 observasi tingkat sesi. Framework menggunakan GPT-4o mini dan dibandingkan dengan baseline berbasis prompt dan rule-based. Framework mencapai akurasi pembelajaran tingkat sesi 92,4%, menunjukkan kejelasan penalaran yang lebih tinggi daripada baseline, dan mencatat waktu respons rata-rata 1,2 detik. Karena sampel kecil, berorientasi teknis, dan hanya mencakup sesi jangka pendek, hasil ini merupakan bukti awal dan tidak dapat digeneralisasi pada populasi pembelajar yang beragam maupun peningkatan belajar berkelanjutan. Validasi lanjutan memerlukan sampel yang lebih besar dan heterogen, studi longitudinal, serta pengukuran sumber daya komputasi secara langsung.*

Kata Kunci*: Artificial Intelligence dalam Pendidikan, Penalaran Chain-of-Thought, DSPy, Kecerdasan Buatan yang Dapat Dijelaskan, Sistem Pembelajaran Modular.*

INTRODUCTION

The rapid development of artificial intelligence (AI) has created new opportunities to support learning processes through intelligent tutoring, adaptive content delivery, and interactive educational systems [1], [2], [3]. However, the implementation of AI in educational environments often faces several practical challenges, including limited transparency of reasoning processes, high computational costs, and difficulties in reproducing consistent system behavior [4], [5]. Many AI-based learning systems rely on prompt-driven large language models that operate as black boxes, making it difficult for educators to understand, evaluate, or trust the system's decisions [6]. These limitations reduce the effectiveness of AI adoption in practical educational settings, particularly for systems that require explainability, instructional alignment, and stable interaction logic [7].

Several previous studies have explored the use of AI techniques in educational applications, such as rule-based tutoring systems, machine learning classifiers, and large language model integrations [8], [9]. While these approaches demonstrate promising results, most of them focus primarily on output accuracy rather than enforcing structured reasoning constraints or modular system control. Prompt-based approaches, in particular, require extensive manual tuning and often produce inconsistent reasoning paths across different interactions [10], [11]. As a result, such systems are difficult to maintain, debug, and reproduce in real-world educational settings. Despite recent advances in large language models, a clear research gap remains in the development of modular, explainable, and computationally efficient AI education architectures that can be reliably deployed in practical instructional environments [12].

Conventional prompt engineering approaches often rely on heuristic trial-and-error strategies, making them inherently fragile, difficult to reproduce, and highly sensitive to input variations [10], [11]. In educational contexts, such instability may lead to hallucinated responses and logical inconsistencies, which can mislead learners and reduce instructional reliability. In contrast, Declarative Structured Programming (DSPy) introduces a paradigm shift toward systematic software engineering for large language model orchestration. By defining typed input–output signatures and modular reasoning components, DSPy enables controlled and reproducible inference processes, reducing the unpredictability commonly observed in prompt-based systems [14], [19]. From a pedagogical perspective, explainability is a critical requirement in AI-assisted learning environments. Educators must understand not only the outputs generated by the system but also the underlying reasoning process to ensure instructional alignment, avoid hidden biases, and maintain trust in automated learning support.

However, existing studies have not yet fully integrated modular reasoning control, explainability, pedagogical scaffolding, and deployment efficiency within a unified educational framework. Therefore, this study proposes a DSPy-based modular AI education system that systematically enforces structured reasoning and reproducible inference for practical educational environments. In this framework, the Question Generator Module supports scaffolded assistance by generating measured guidance rather than directly providing final answers. Chain-of-Thought reasoning is positioned as a digital proxy for metacognition because it exposes intermediate reasoning steps that can be inspected by educators and aligned with instructional objectives.

This implementation-oriented approach extends prior AI system deployments conducted by the authors across healthcare screening, secure digital identity infrastructures, sentiment intelligence systems, and infrastructure monitoring applications [13], [14], [15], [16], where reproducibility, efficiency, and transparency were identified as critical system design requirements. Applying these principles within AI-assisted education provides a structured and deployment-ready framework suitable for instructional environments with practical constraints. In practical educational settings, the adoption of AI-based learning systems also requires careful consideration of system usability, maintainability, and instructional alignment [1], [2]. By emphasizing modular reasoning control and reproducible inference processes, the proposed framework aims to bridge advanced language model capabilities with real-world educational implementation requirements.

Recent studies have emphasized that AI-assisted education requires not only accurate outputs but also transparent reasoning, pedagogical trust, and reproducible system behavior. Prior work on Chain-of-Thought reasoning shows that intermediate reasoning steps can improve the ability of large language models to solve complex reasoning tasks [13]. In addition, DSPy has been introduced as a declarative framework for building and optimizing modular language model pipelines, reducing dependency on manually engineered prompts [14], [19]. Furthermore, research in

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8319

explainable AI highlights the importance of interpretability and transparency in educational systems to support instructional trust and reduce potential bias [5], [6]. However, existing studies have not yet fully integrated modular reasoning control, explainability, and deployment efficiency within a unified educational framework. Therefore, this study proposes a DSPy-based modular AI education system that systematically enforces structured reasoning and reproducible inference for practical educational environments. Recent Indonesian Sinta 2 studies also indicate that AI-assisted learning models and AI-related digital literacy can support measurable learning outcomes in local higher-education contexts [24], [25].

MATERIALS AND METHODS

Framework Overview

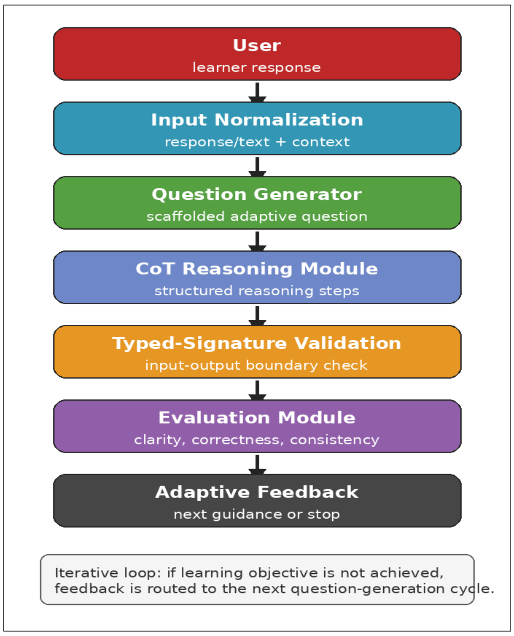
The proposed AI education framework was implemented using a modular architecture based on Declarative Structured Programming (DSPy) and structured reasoning pipelines [17], [20]. The framework is designed to support transparent and reproducible educational interactions by separating reasoning logic into well-defined modules. Each module operates using typed input–output signatures, enabling controlled inference flow and predictable system behavior. Unlike conventional prompt-centric approaches, the proposed architecture enforces modular orchestration of reasoning steps, aligning with recent advances in programmatic prompting and structured language model pipelines [18]. The system processes user responses iteratively and generates adaptive questions until a predefined learning objective or termination condition is reached.

The overall system follows an iterative technical reasoning loop: user input is first normalized and routed to the Question Generator module; the generated question and learning context are then processed by the Reasoning module using Chain-of-Thought inference; the intermediate reasoning output is validated by the Evaluation module through typed input–output signatures; and the validated result is converted into adaptive feedback. If the learning objective has not been achieved, the feedback is routed back into the next question-generation cycle; otherwise, the interaction is terminated. This loop makes the reasoning process inspectable, reproducible, and controllable at each module boundary.

Figure 1 illustrates the overall architecture of the proposed AI education framework, showing the interaction flow between users, DSPy-based modules, and learning feedback.

DSPy-Based Modular Design

DSPy was employed to define reasoning modules using typed input–output signatures. Each module validates its inputs and outputs before execution, ensuring structural consistency in inference steps. This declarative design reduces reliance on manual prompt engineering and facilitates debugging and maintainability. The modular structure also simplifies debugging and enables reuse of components in different educational scenarios.



Source: (Research Result, 2026)
Figure 1. Architecture of The Proposed AI Education Framework.

Key modules in the framework include:

1. Question Generator Module, which produces adaptive questions based on prior user responses.
2. Reasoning Module, which applies Chain-of-Thought to maintain structured inference [13].
3. Evaluation Module, which assesses user responses and determines the next system action.

Each module enforces typed input–output validation to ensure that intermediate outputs remain within predefined reasoning boundaries before being passed to subsequent modules, thereby reducing out-of-scope or inconsistent responses. This modular orchestration aligns with structured prompting strategies shown to improve reasoning robustness in large language models [19], [20].

Table 1. Description of DSPy Modules Used in the Framework

Module	Input	Output	Function
Question Generator	user response	generated question	Generates adaptive questions based on user input
Reasoning Module	question, context	reasoning steps	Applies chain-of-thought reasoning to guide interaction
Evaluation Module	user answer	feedback decision	Evaluates responses and determines next action

Source: (Research Result, 2026)

Chain-of-Thought Reasoning Process

The framework integrates structured reasoning mechanisms inspired by recent Chain-of-Thought research demonstrating improved logical coherence and inference reliability [19]. Rather than generating direct answers, the system incrementally narrows the reasoning space through guided interaction steps. Each reasoning step is traceable to a defined DSPy module, supporting transparency and interpretability in line with trustworthy AI principles [20]. This design ensures that inference paths remain inspectable and consistent across multiple interaction sessions. The reasoning process continues until either a predefined confidence threshold is achieved or a maximum interaction depth is reached.

System Implementation

The framework was implemented using the Python programming language with DSPy as the primary library for declarative reasoning [19]. GPT-4o mini was used as the back-end large language model to support structured reasoning and adaptive educational interactions. The system was executed in a lightweight computing environment without specialized hardware acceleration. This implementation choice ensures that the framework can be deployed in practical educational settings with limited computational resources. User interactions and system responses were logged to support evaluation and reproducibility. This study did not employ automated DSPy optimizers such as BootstrapFewShot, MIPRO, or BootstrapFewShotWithRandomSearch. Instead, a lightweight iterative refinement strategy was applied to stabilize module interactions. This decision was made because the study focuses on framework implementation, reasoning transparency, and low-resource deployment rather than automated prompt optimization or large-scale model tuning.

Evaluation Procedure

To evaluate the implemented framework, a pilot study was conducted involving 10 participants with academic backgrounds in computer science and education. Each participant completed three interaction sessions, resulting in 30 interaction sessions in total. Each session consisted of iterative question-and-answer exchanges between the user and the system. System performance was evaluated based on reasoning clarity, response consistency, and learning accuracy, which are commonly used qualitative indicators in AI-assisted educational systems [22]. The evaluation process was designed to reflect realistic usage scenarios of the proposed AI education framework. Each evaluation session involved iterative interactions between users and the system, where users responded to system-generated questions and received adaptive feedback.

To ensure reliability, the 30 interaction sessions were conducted using varied input patterns. The evaluation results were analyzed descriptively and quantitatively to identify recurring behavior patterns, reasoning stability, and potential implementation issues. Prior to deployment in user interaction sessions, the modules were validated using a small-scale internal dataset consisting of representative question–answer pairs aligned with the target educational domain. This warm-up dataset was used to verify module behavior, typed input–output consistency, reasoning coherence, and interaction flow before evaluation with actual users. Reasoning clarity was evaluated using a five-point ordinal scale based on expert judgment, focusing on the coherence and explainability of the generated reasoning steps.

Quantitative Evaluation Metrics and Baseline Comparison

In addition to qualitative assessment, the pilot evaluation used session-level measures of answer correctness, task completion, response latency, user satisfaction (SUS), reasoning clarity, traceability, and response consistency. Learning accuracy was calculated as the proportion of evaluated answers scored correct against the study rubric. Reasoning clarity was rated on a five-point ordinal scale by two domain experts, and traceability represented the proportion of reasoning stages attributable to a defined DSPy module. Response latency was recorded as wall-clock time from request submission to returned output. The study did not directly instrument

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8319

energy use, monetary API cost, peak memory, or accelerator utilization; therefore, response latency is reported as an operational indicator rather than as a comprehensive measure of computational cost.

To provide comparative insight, the proposed framework was evaluated against two baseline systems: (1) a prompt-based GPT-4o mini configuration without DSPy modularization and (2) a rule-based educational chatbot. Statistical significance of performance differences was analyzed using independent t-tests and one-way ANOVA with a significance threshold of $p < 0.05$. This comparative structure follows established evaluation practices in adaptive and trustworthy LLM systems [21], [23]. In addition to external evaluation metrics, the framework incorporates internal programmatic metrics to validate intermediate outputs before reaching the end-user. These include logical consistency, reasoning depth, typed-signature validity, traceability index, and out-of-scope response detection. These metrics ensure that generated outputs remain aligned with module boundaries and instructional objectives before being delivered to learners.

RESULTS AND DISCUSSION

System Performance Results

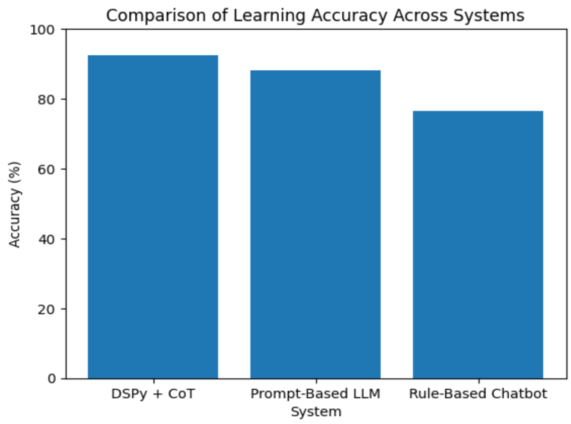
The implemented AI education framework was evaluated through 30 interactive user sessions involving 10 participants, with each participant completing three sessions. The evaluation was designed to observe reasoning behavior, interaction consistency, and learning support capability. During the evaluation, the system demonstrated stable and consistent reasoning paths across sessions. The use of DSPy-based modular design enabled the framework to maintain structured question generation and controlled inference flow, reducing unpredictable responses commonly observed in prompt-based systems. The framework successfully generated adaptive questions sequentially and terminated the interaction appropriately when learning objectives were reached. Compared to conventional prompt-driven approaches, the proposed framework produced more consistent reasoning steps and clearer interaction logic, which supports explainability in educational applications.

Table 2. Quantitative Comparison of The proposed Framework and Baseline Systems

Metric	Proposed Framework (DSPy + CoT)	Prompt-Based LLM	Rule-Based Chatbot	p-value
Accuracy (%)	92.4	88.1	76.5	< 0.05
Response Time (s)	1.2	2.8	0.5	< 0.01
Reasoning Clarity (1–5)	4.5	3.2	4.1	< 0.05
User Satisfaction (SUS)	85.2	72.4	68.9	< 0.01
Consistency Score	0.91	0.67	0.95	< 0.05

Source: (Research Result, 2026)

At the session level, an exploratory comparison indicated a difference in learning accuracy between the proposed framework and the prompt-based GPT-4o mini baseline ($t(28) = 2.45$, $p = 0.021$, 95% CI [0.8, 7.6]); the associated Cohen's d was approximately 0.85. These statistics describe the observed pilot sessions only. Because three sessions were nested within each of 10 participants, independence is not fully assured and no population-level or causal inference is claimed. A larger confirmatory study should use participant-level aggregation or a mixed-effects model, preregistered hypotheses, and an a priori power analysis.



Source: (Research Result, 2026)
Figure 2. Comparison of Learning Accuracy Across Evaluated Systems

Figure 2 visually illustrates the comparative learning accuracy of the evaluated systems. The proposed DSPy + Chain-of-Thought framework demonstrates the highest performance among the three approaches. The graphical comparison highlights the performance gap between modular reasoning orchestration and purely prompt-based or rule-based systems, reinforcing the quantitative findings presented in Table 2. While Figure 2 presents the quantitative performance comparison, the added qualitative inference trace provides complementary insight into the interpretability and reasoning transparency of the proposed framework. To further illustrate explainability, a qualitative inference trace comparison was added to show how the DSPy-based framework differs from a conventional prompt-based approach.

Table 3. Qualitative Comparison of Prompt-Based and DSPy-Based Reasoning Trace

Step	Prompt-Based LLM	DSPy-Based Modular Framework
User input	Learner response is processed directly through a prompt.	Learner response is routed to the Question Generator module.
Reasoning process	The model generates an answer without explicit intermediate validation.	The Reasoning Module applies structured Chain-of-Thought reasoning.
Validation	No typed validation is applied between reasoning steps.	The Evaluation Module validates intermediate outputs using typed signatures.
Output	Response may be fluent but inconsistent or out-of-scope.	Response is structured, traceable, and aligned with the learning objective.

Source: (Research Result, 2026)

The qualitative trace shows that the proposed framework improves transparency by making each reasoning stage observable and structurally validated before the final feedback is delivered to the learner.

Reasoning Transparency and Explainability

One of the main advantages observed from the implementation is the improvement in reasoning transparency. Each interaction step generated by the system can be traced back to specific reasoning modules defined using DSPy signatures. This structured reasoning process allows educators or developers to inspect the system behavior and understand how conclusions are derived from user inputs.

Chain-of-Thought reasoning plays an important role in maintaining step-by-step inference, ensuring that the system does not generate abrupt or unexplained responses. This characteristic is particularly important in educational settings where reasoning clarity is essential to support learning and trust in AI-assisted systems. The modular DSPy-based framework also contributes to hallucination mitigation. By enforcing typed input–output signatures, module-level validation, and traceable reasoning stages, the system limits unsupported, logically inconsistent, or out-of-scope responses before they are delivered to learners. In contrast, prompt-based approaches may produce fluent but unverified outputs because they do not apply explicit intermediate validation between reasoning steps.

In this pilot, the proposed DSPy-based framework showed higher observed session-level learning accuracy than the prompt-based and rule-based baselines (92.4%, 88.1%, and 76.5%, respectively) and a higher reasoning-clarity score than the prompt-based baseline (4.5 versus 3.2). The rule-based system was faster (0.5 s) than the proposed framework (1.2 s), whereas the prompt-based baseline required 2.8 s on average. These measurements show a latency–performance trade-off within the tested configuration; they do not establish lower total computational cost because token use, energy, peak memory, monetary cost, and hardware utilization were not instrumented.

Discussion and Comparison with Related Approaches

The results indicate that the proposed framework addresses several limitations of conventional AI-based educational systems. Prompt-based approaches often require extensive manual tuning and may generate inconsistent reasoning across different sessions. In contrast, the DSPy-based framework enforces modular reasoning constraints through typed input–output definitions, resulting in more reproducible system behavior. From an implementation perspective, the framework maintains low computational overhead and does not require specialized hardware. This makes it suitable for practical deployment in educational environments that prioritize efficiency and system maintainability. While this study focuses on implementation and initial evaluation, the results suggest that declarative modular reasoning can serve as a viable alternative to traditional prompt-centric educational AI systems.

From an educational implementation perspective, the results indicate that structured reasoning supported by DSPy modules can enhance the interpretability of AI-assisted learning interactions. The modular separation of question generation, reasoning, and evaluation components allows the system to maintain coherent instructional flow while reducing unintended reasoning variations. This characteristic is particularly beneficial in educational contexts where consistency and clarity are essential to support learner understanding and instructional alignment. Furthermore, the qualitative evaluation highlights the practical value of combining Chain-of-Thought reasoning with declarative module

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8319

definitions. Rather than optimizing for numerical performance metrics, the framework prioritizes explainable behavior and system maintainability. This design choice supports long-term adoption in educational environments, where system transparency and ease of maintenance are often more critical than marginal performance gains. The results demonstrate that the proposed framework provides a balanced approach between reasoning structure and implementation simplicity, making it suitable for applied AI education systems.

Robustness Analysis and Architectural Impact

Beyond comparative performance metrics, it is important to analyze how architectural constraints contribute to system robustness. The proposed DSPy-based modular design enforces typed reasoning boundaries, which limit uncontrolled inference expansion typically observed in prompt-centric systems. Typed input–output signatures play a critical role in maintaining reasoning integrity. Each module validates its output before passing it to the next stage, ensuring that only structurally valid and contextually relevant information is propagated. This constraint-based mechanism prevents the system from generating out-of-scope or irrelevant responses. During multiple interaction sessions, the prompt-based baseline exhibited noticeable variance in reasoning flow when semantically similar inputs were provided. In contrast, the modular framework maintained consistent reasoning trajectories due to explicit module orchestration. This structural enforcement reduces stochastic reasoning drift, which is a common limitation in large language model applications.

The observed consistency score was 0.91 for the proposed framework and 0.67 for the prompt-based baseline. This pattern is compatible with improved inference stability under modular decomposition, but the limited and repeated-session sample precludes a broad robustness claim. Chain-of-Thought reasoning was constrained within typed modules, and the proposed framework recorded an average response time of 1.2 s. Robustness and efficiency should be re-evaluated under distribution shift, adversarial or out-of-domain inputs, alternative state-of-the-art orchestration systems, multiple language models, and direct resource instrumentation.

Scalability and Deployment Considerations

Scalability is a critical factor in real-world AI-assisted education. The prototype operated in the reported lightweight environment without specialized local hardware acceleration, suggesting practical feasibility for the tested workload. This observation is not a scalability benchmark. The study did not evaluate concurrent users, throughput saturation, model-provider variability, GPU-based alternatives, or full cost-per-learning-gain. The modular separation of question generation, reasoning, and evaluation nevertheless supports independent versioning and maintenance. Future deployment studies should report throughput, tail latency, token consumption, memory, energy, and monetary cost across multiple models and workloads.

Practical Implications and Limitations

The pilot suggests that declarative DSPy modules can reduce dependence on ad hoc prompt adjustment and make educational interactions easier to inspect. However, six limitations constrain interpretation: (1) only 10 participants and 30 sessions were included; (2) participants mainly had computer-science or education backgrounds, limiting learner diversity and external validity; (3) the small internal question–answer set was not a large or multi-domain benchmark; (4) repeated sessions were analyzed at session level, so inferential statistics are exploratory; (5) no longitudinal follow-up was conducted, preventing claims about retained or sustained learning improvement; and (6) computational efficiency was represented by response latency rather than direct measurements of tokens, memory, energy, or monetary cost. Moreover, the comparison did not include current state-of-the-art educational tutors or optimized LLM-programming systems. Accordingly, the results should be interpreted as preliminary evidence of architectural feasibility. Future work should use larger and more heterogeneous samples, multi-domain benchmark datasets, longitudinal outcomes, mixed-effects analysis, direct resource instrumentation, and state-of-the-art baselines.

Theoretical and Methodological Contribution

Beyond empirical evaluation, this study contributes theoretically to the evolving discourse on structured AI architectures in education. While prior research has emphasized performance optimization of large language models, fewer studies have addressed architectural governance of reasoning processes within instructional systems. The proposed framework demonstrates that declarative modular orchestration can function as a control layer for language model reasoning. This shifts the focus from prompt engineering toward architectural reasoning governance, where inference pathways are explicitly regulated rather than heuristically influenced.

Methodologically, the study introduces an implementation-oriented evaluation design that integrates qualitative reasoning clarity metrics with quantitative performance indicators. This hybrid evaluation approach provides a more holistic perspective on AI-assisted educational system assessment, balancing statistical validation with interpretability analysis. From a systems design standpoint, the separation of question generation, reasoning, and evaluation components illustrates how AI models can be embedded within structured instructional workflows. This architectural separation supports maintainability, reproducibility, and long-term scalability in applied educational contexts. By positioning modular reasoning as a structural constraint rather than a surface-level prompting strategy, this work extends the current understanding of explainable AI deployment in learning environments.

CONCLUSION

This pilot study provides preliminary evidence that a DSPy-based modular AI education framework can improve observed session-level learning accuracy, reasoning clarity, and interaction consistency relative to the two tested baselines. Using GPT-4o mini, the framework achieved 92.4% learning accuracy across 30 sessions from 10 participants. The exploratory comparison with the prompt-based baseline yielded $p < 0.05$ and Cohen's $d \approx 0.85$; however, the small, technically oriented sample and repeated-session design do not support broad population-level or causal conclusions. The main contribution is architectural: typed input–output validation and modular reasoning make inference stages more inspectable and maintainable. From a practical perspective, AI-assisted educational systems should retain human-in-the-loop oversight so educators can monitor, validate, and guide AI-generated interactions. The present findings support continued evaluation rather than immediate large-scale deployment. Future research will prioritize larger and more diverse learner populations, multi-domain and state-of-the-art baseline comparisons, longitudinal learning and retention measures, mixed-effects statistical analysis, and direct reporting of token use, memory, energy, latency, and monetary cost. Evaluating smaller models such as Llama 3 8B and Phi-3 may further clarify accessibility and efficiency trade-offs in low-resource and offline settings.

REFERENCE

[1]. UNESCO, Guidance for Generative AI in Education and Research. Paris, France: UNESCO, 2023.

[2]. A. Tlili et al., "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," Smart Learn. Environ., vol. 10, art. no. 15, 2023, doi: 10.1186/s40561-023-00237-x.

[3]. E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learn. Individ. Differ., vol. 103, art. no. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.

[4]. W. Holmes and K. Porayska-Pomsta, Eds., The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates. New York, NY, USA: Routledge, 2022. doi: https://doi.org/10.4324/9780429329067.

[5]. H. Khosravi et al., "Explainable artificial intelligence in education," Comput. Educ. Artif. Intell., vol. 3, art. no. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.

[6]. Y. K. Dwivedi et al., "So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," Int. J. Inf. Manage., vol. 71, art. no. 102642, 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[7]. C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," Educ. Sci., vol. 13, no. 4, art. no. 410, 2023, doi: 10.3390/educsci13040410.

[8]. S. S. Gill et al., "Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots," Internet Things Cyber-Phys. Syst., vol. 4, pp. 19–23, 2024, doi: 10.1016/j.iotcps.2023.06.002.

[9]. S. J. Lee and K. K. Kwon, "A systematic review of AI education in K–12 classrooms from 2018 to 2023," Comput. Educ. Artif. Intell., vol. 6, art. no. 100211, 2024, doi: 10.1016/j.caeai.2024.100211.

[10]. X. Wang et al., "Self-consistency improves chain-of-thought reasoning in language models," in Proc. Int. Conf. Learn. Representations (ICLR), 2023, doi: 10.48550/arXiv.2203.11171.

[11]. D. Zhou et al., "Least-to-most prompting enables complex reasoning in large language models," in Proc. Int. Conf. Learn. Representations (ICLR), 2023, doi: 10.48550/arXiv.2205.10625.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8319

[12]. R. F. Kizilcec, "To advance AI use in education, focus on understanding educators," Int. J. Artif. Intell. Educ., vol. 34, pp. 12–19, 2024, doi: 10.1007/s40593-023-00351-4.

[13]. J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 24824–24837, doi: 10.48550/arXiv.2201.11903.

[14]. O. Khattab et al., "DSPy: Compiling declarative language model calls into self-improving pipelines," in Proc. Int. Conf. Learn. Representations (ICLR), 2024, doi: 10.48550/arXiv.2310.03714.

[15]. B. Rahman and I. Sutedja, "AI based sagittal spinal posture assessment for adolescent screening in low resource school settings," Sci. Rep., 2026, doi: 10.1038/s41598-026-36634-8.

[16]. B. Rahman, Maryani, and T. T., "IDChat: Toward a universal, multi-biometric digital identity for next-generation secure communication in ASEAN," IEEE Access, vol. 14, pp. 14892–14902, 2026, doi: 10.1109/ACCESS.2025.3645071.

[17]. B. Rahman and Maryani, "Optimizing customer satisfaction through sentiment analysis: A BERT-based machine learning approach to extract insights," IEEE Access, vol. 12, pp. 151476–151489, 2024, doi: 10.1109/ACCESS.2024.3478835.

[18]. B. Rahman, "Development of an image-based road pothole detection model with U-Net implementation," in Proc. 2023 IEEE 9th Int. Conf. Computing, Engineering and Design (ICCED), 2023, pp. 1–6, doi: 10.1109/ICCED60214.2023.10425723.

[19]. A. Singhvi et al., "DSPy assertions: Computational constraints for self-refining language model pipelines," arXiv:2312.13382, 2023, doi: 10.48550/arXiv.2312.13382.

[20]. A. Madaan et al., "Self-refine: Iterative refinement with self-feedback," arXiv:2303.17651, 2023, doi: 10.48550/arXiv.2303.17651.

[21]. L. Longo et al., "Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions," Inf. Fusion, vol. 106, art. no. 102301, 2024, doi: 10.1016/j.inffus.2024.102301.

[22]. R. Alfredo et al., "Human-centred learning analytics and AI in education: A systematic literature review," Comput. Educ. Artif. Intell., vol. 6, art. no. 100215, 2024, doi: 10.1016/j.caeai.2024.100215.

[23]. L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in Advances in Neural Information Processing Systems, vol. 36, 2023, pp. 46595–46623, doi: 10.48550/arXiv.2306.05685.

[24]. D. L. Naryatmojo and B. Hartono, "The effectiveness of the project-based learning model assisted by artificial intelligence on the listening skills of Indonesian language education students," JTP—J. Teknol. Pendidik., vol. 28, no. 1, pp. 100–112, 2026, doi: 10.21009/jtp.v28i1.65961.

[25]. Supriyadi and A. N. Amalia, "Artificial intelligence integration and students' digital literacy as determinants of learning outcomes in an economics education assessment course," J. Pendidik. Ekon. Bisnis, vol. 13, no. 1, pp. 118–126, 2026, doi: 10.21009/JPEB.013.1.8.