Open Access

Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact and effectiveness of personalized feedback

William Villegas-Ch[1*] , Diego Buenano-Fernandez[1†], Alexandra Maldonado Navarro[2†] and Aracely Mera-Navarrete[3†]

†Diego Buenano-Fernandez, Alexandra Maldonado Navarro and Aracely Mera-Navarrete have contributed equally to this work.

*Correspondence:
william.villegas@udla.edu.ec

[1] Escuela de Ingeniería en Ciberseguridad, Universidad de Las Américas, 170125 Quito, Ecuador
[2] Maestría en Derecho Digital, Facultad de postgrados, Universidad de Las Américas, 170125 Quito, Ecuador
[3] Departamento de Sistemas, Universidad Internacional del Ecuador, 170411 Quito, Ecuador

Abstract

Teaching Science, Technology, Engineering, and Mathematics (STEM) disciplines faces significant challenges due to the increasing complexity of content and the diversity in students' knowledge levels. Traditional teaching methods, characterized by their standardized approach, fail to adapt to individual needs, limiting learning potential. This problem is particularly evident in mathematics, physics, and programming, where mastering critical concepts requires a logical and sequential progression. This study proposes an artificial intelligence-based intelligent tutoring system to provide a real-time personalized and adaptive learning experience. The system integrates advanced deep learning and natural language processing models, allowing for targeted feedback and dynamic adjustment of learning trajectories. The results show significant improvements in several key metrics. The experimental group achieved an average precision of 85% in programming and 78% in mathematics, significantly outperforming the control group. Furthermore, a linear regression model identified a positive correlation between the time of interaction with the system and the rate of progress in fundamental concepts. Student perceptions were also highly positive, with 80% appreciating the usefulness of adaptive feedback. These findings identify the system's potential to transform STEM teaching, address the lack of personalization, and improve learning in various educational settings.

Keywords: Intelligent tutoring systems, STEM education, Adaptive learning, Artificial intelligence in education

Introduction

Teaching Science, Technology, Engineering, and Mathematics (STEM) disciplines face significant challenges today due to the increasing complexity of content, heterogeneity in students' knowledge levels, and the need to adapt pedagogical methods to digital contexts (Chen et al., 2024). These challenges are particularly evident in mathematics, physics, and programming, where mastery of critical concepts depends on a logical and sequential progression. Traditionally, teaching these disciplines has employed

Villegas-Ch *et al. Smart Learning Environments* (2025) 12:41

Page 2 of 31

standardized approaches that, while effective for certain groups, fail to accommodate individual needs, thereby limiting the learning potential of many students (Dlouhy and Froidevaux, 2024).

Intelligent tutoring systems (ITS) have emerged as a promising solution (Spitzer and Moeller, 2023). These systems, enabled by advances in artificial intelligence (AI), offer capabilities to personalize learning through real-time analysis of student performance and adaptive content recommendations. Previous research has demonstrated the effectiveness of ITS in improving academic performance, especially in highly structured domains such as mathematics and programming (Muangprathub et al., 2020). However, these studies have also highlighted critical limitations, such as the lack of real-time dynamic adaptability and the reliance on probabilistic models that fail to capture the complexity of the data generated by students.

Our study addresses these limitations by developing and implementing an intelligent tutoring system for STEM disciplines, designed with a modular architecture that integrates deep learning models and natural language processing (NLP) techniques (Jabbar et al., 2024). This system assesses student progress in real time, dynamically adjusts learning trajectories, and provides targeted feedback to address complex concepts. The methodology included evaluating 450 university students, divided into experimental and control groups, to analyze the system's precision, rate of progress, and perceived progress in mathematics, physics, and programming.

The adaptive approach of the system responds to the urgent need for effective and inclusive teaching methodologies. Digital technologies have transformed education by enabling the collection of massive amounts of data on students' learning. However, most educational environments still lack the tools to leverage this potential to personalize the educational experience (Zhu, 2024). Our system seeks to bridge this gap by combining advanced AI techniques with sound pedagogical principles. This system addresses fundamental problems, including a lack of effective personalization, reliance on static feedback, and limitations in progress assessment. Unlike traditional approaches, our system uses transformer-based models to analyze student interactions and provide targeted recommendations. Furthermore, it adapts feedback based on historical and real-time performance, significantly improving precision in solving exercises (Han, 2024).

The developed system was evaluated in a controlled environment over one academic semester. The system architecture included modules for natural language processing, content recommendation, and real-time assessment. Students in the experimental group interacted with the system to solve exercises and receive adaptive feedback, whereas the control group used traditional, static, resource-assisted teaching methods. Data on precision, progress rate, and student perception were collected to measure the impact. A linear regression model was used to analyze the relationship between interaction time and progress on critical concepts, revealing a positive correlation with a coefficient of determination ($R^2 = 0.76$). Furthermore, statistical results showed significant differences between the two groups in terms of precision ($p = 0.002$) and perception of progress ($p = 0.032$), confirming the effectiveness of the system.

The quantitative and qualitative results highlight the system's effectiveness in multiple dimensions. In mathematics, students in the experimental group mastered 78% of the assessed concepts, while in physics and programming, they achieved rates of 70% and

85%, respectively. Students with high interaction achieved progress rates above 90% in programming and 80% in physics, demonstrating a direct impact of continuous interaction. The adaptive feedback provided was highly valued, with 80% of students considering it helpful in identifying and correcting errors. These results reinforce the hypothesis that an adaptive and personalized design can overcome the limitations of traditional methods.

The developed system represents a significant advance in the application of AI technologies in STEM education. Its ability to integrate real-time analysis, adaptive feedback, and a scalable modular architecture positions it as an innovative tool to overcome the limitations of traditional methods and other existing ITS (Ong et al., 2023). Furthermore, by combining rigorous statistical analysis with ethical principles in data management, this work sets a standard for future developments in the field. Despite the promising results, this study also presents limitations. The reliance on a homogeneous population of university students restricts the generalization of the findings to other educational contexts (Yeung et al., 2024).

Furthermore, the lack of validation in fully real-world scenarios introduces uncertainty about the system's applicability to populations with significantly different characteristics. These restrictions underscore the need for future research to validate the results in more diverse populations and investigate their applicability to other unstructured disciplines (Beckett et al., 2024). This work highlights the significance of AI-based ITS in transforming STEM teaching, laying a solid foundation for future research on enhancing and expanding its educational impact.

Literature review

ITS has proven effective in structured disciplines such as mathematics and programming, where concepts can be modeled sequentially and hierarchically. For example, Kolekar et al. (2019) analyzed the impact of rule-based ITS and concluded that these systems achieve an average improvement equivalent to one standard deviation over traditional methods. However, their work highlights limitations in real-time adaptive capability, where modern systems based on neural networks and deep learning have shown significant advances.

Personalization of learning is a critical feature in modern ITS. Nimy et al. (2023) proposed a probabilistic model to predict students' knowledge levels based on interaction data, achieving effective personalization in algebra topics. Although this approach proved promising, the reliance on static models limited its applicability in dynamic domains, such as programming, where students can approach problems from multiple valid approaches. This limitation has been addressed by recent work integrating transformer-based architectures, as seen in the study by Tian et al. (2023), which utilized pre-trained language models to generate adaptive feedback on coding problems, thereby significantly improving learning speed.

In automatic feedback, several studies have investigated how to provide targeted and valuable suggestions to enhance performance. Turan and Yilmaz (2024) focused on feedback in massive online courses (MOOCs), using clustering algorithms to identify common patterns in student errors. This approach stood out for its scalability. However, its implementation in more controlled environments, such as educational laboratories, has

been limited due to the lack of specific customization for each student. On the other hand, Descalço et al. (2018) introduced Bayesian learning models to track conceptual progress, highlighting their effectiveness in predicting performance in consecutive exercises. Although this approach is robust, its implementation in real-time systems still faces computational challenges.

The relationship between interaction time and learning has been widely explored. A relevant study is that of Septian et al. (2021), who identified a positive correlation between time spent solving problems and concept retention rate in science courses. However, their analysis was limited to homogeneous populations, leaving a gap in exploring how different initial skill levels affect this relationship. In contrast, our work addresses this gap by analyzing heterogeneous populations with varying interaction levels and initial STEM skills.

Ethical and privacy assessments in STI have gained relevance with the use of big data. Studies such as Dari et al. (2024) have emphasized the need for data anonymization strategies and the protection of sensitive information, highlighting that a lack of trust in data management can limit the adoption of these technologies. Our system integrates these recommendations through encryption and anonymization techniques, ensuring that the implementation is both ethical and compliant with international standards.

Identified research gap and justification

While previous studies have contributed significantly to the development of ITS, several limitations persist that justify the need for this research. Rule-based approaches (Kolekar et al., 2019) have shown improvements in learner outcomes but often lack the flexibility required for dynamic, real-time adaptation. Probabilistic models (Nimy et al., 2023) address uncertainty and provide some personalization, yet they are typically static and struggle to generalize in highly interactive environments such as programming tasks. Recent advances in language models (Tian et al., 2023) and feedback mechanisms (Descalço et al., 2018; Turan and Yilmaz, 2024) have focused on single domains or platforms (e.g., MOOCs), but there is limited evidence of their integration into modular and adaptive systems applicable across diverse STEM disciplines.

Moreover, few studies combine semantic feedback, content adaptation, and real-time analysis in a unified architecture, as most solutions isolate these functionalities. While the correlation between interaction time and learning outcomes has been acknowledged in platforms such as Google Classroom (Septian et al., 2021), this relationship has rarely been translated into adaptive instructional strategies within ITS frameworks capable of adjusting content and feedback dynamically based on user engagement patterns. Finally, although ethical considerations are emerging (Dari et al., 2024), practical implementations with privacy-preserving mechanisms are still scarce.

This study addresses these gaps by proposing an integrated ITS architecture that unifies adaptive feedback, natural language processing, personalized content recommendation, and real-time assessment across multiple STEM areas. Additionally, it evaluates the system's performance in a heterogeneous educational setting, using quantitative and qualitative methods to validate its effectiveness, while embedding ethical safeguards to ensure data protection. This responds directly to the existing literature's identified limitations in scope, adaptability, and operational integration.

Materials and methods

Architecture of the smart tutoring system

The architecture of the intelligent tutoring system for teaching STEM skills is designed to provide an adaptive and personalized learning experience based on a modular infrastructure that enables seamless interaction between various AI components (Ouyang and Xu, 2024). The system's structure is organized into specific modules, each with a distinct functionality that contributes to personalizing and adapting content, understanding student queries, recommending materials, and providing real-time assessment of learning progress. This modular design facilitates scalability and flexibility, allowing the system to adapt to various disciplines within STEM and different levels of knowledge (Hurley et al., 2024).

General architecture description

The smart tutoring system is based on a distributed architecture composed of several interdependent modules, including the natural language processing (NLP) module, the content recommendation module, and the real-time assessment module (Rathi et al., 2023). Each module is managed by a central control layer that coordinates the communication between the components and facilitates the integration of AI models in an efficient and scalable manner (Aguayo et al., 2021). Module communication is managed through an application programming interface (API), which enables real-time data transfer and ensures information synchronization regarding the student's profile, progress, and needs.

Figure 1 presents the architecture of the intelligent tutoring system for teaching STEM skills, highlighting the modular arrangement and the data flow between the main components. This diagram allows us to visualize how the NLP, content recommendation, and real-time assessment modules interact and coordinate their functions through an API, ensuring an adaptive and personalized learning experience.

This architecture's foundation is a distributed storage database that keeps historical records of student interactions, including responses, error patterns, and response times. This database is crucial for the intelligent tutoring system to tailor its
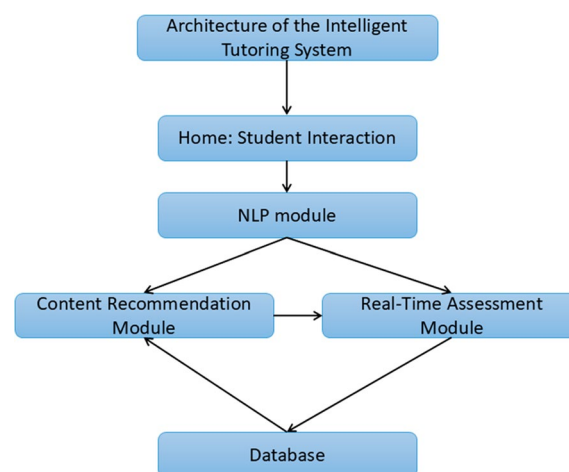

Fig. 1 Modular architecture of the smart tutoring system for teaching STEM skills

Villegas-Ch *et al. Smart Learning Environments*   (2025) 12:41

Page 6 of 31

recommendations and feedback to the student's past behavior. The data flow between modules is continuous, ensuring the system can process information in real-time and adjust its behavior as the student progresses.

Figure 2 illustrates a representative view of the system's user interface to complement the modular architecture described. This interface enables students to engage with personalized exercises, receive real-time feedback, and navigate adaptive learning paths tailored to their performance and preferences.

Natural language processing (NLP) module

The NLP module interprets and analyzes student queries and answers, providing semantic understanding that enables the system to interact with the user effectively. This module utilizes advanced neural network architectures, including transformers and recurrent neural networks (RNNs), which are optimized to capture contextual relationships and syntactic patterns in natural language (Alshawi et al., 2024). The transformer architecture, specifically attention models such as Bidirectional Encoder Representations from Transformers (BERT), enables the analysis of words and phrases in a bidirectional context, optimizing the understanding of questions asked by students and generating accurate answers based on the query's intent (Subakti et al., 2022).

This module comprises two fundamental subcomponents: a semantic analyzer and an answer generator. The semantic analyzer converts queries into feature vectors using embeddings generated by the transformer model, allowing it to identify key terms and detect relationships between concepts in STEM. The response generator uses these vector representations to build responses that integrate the system's knowledge and adjust the level of complexity according to the student's profile. In addition, the NLP module is complemented by a disambiguation component, which refines queries in case of confusing or poorly formulated responses, thus improving the system's precision in the tutoring process (Rosenzweig et al., 2024).
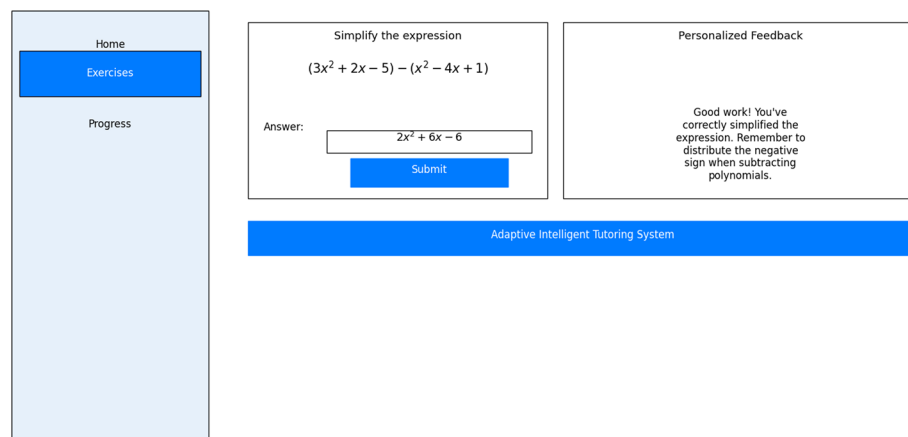

Fig. 2 Representative Interface of the Intelligent Tutoring System showing personalized feedback, interactive exercises, and adaptive navigation

Content recommendation module

The content recommendation module personalizes the proposed learning material and activities based on each student's profile and progress history. This module utilizes collaborative filtering algorithms and recommendation techniques based on hybrid systems, which combine collaborative and content-based filtering approaches. Through the analysis of interaction patterns and the specific interests of each user, this module identifies areas of interest or difficulty for each user. It suggests content that aligns with the individual learning level and needs in STEM.

The architecture of this module is based on a hybrid system that combines recommendations derived from user similarity with those based on content similarity (Sahin et al., 2024). The data collected in the database includes academic performance in specific exercises and the user's preferences expressed in STEM-related topics, enabling precise personalization. This approach ensures that the system suggests additional resources in areas where the student performs poorly, improving the learning process through adapted and relevant content. The recommendation engine also updates in real-time, adjusting recommendations as the learner interacts with the content, maximizing the relevance of suggested materials at every stage of the learning process.

Real-time evaluation module

The real-time assessment module is at the core of the tutoring system's adaptability, as it enables the adjustment of difficulty level and content based on the student's performance in each interaction. This module utilizes pattern analysis algorithms and machine learning techniques to evaluate student progress, taking into account metrics such as response time, response accuracy, and recurring errors. Continuous assessment allows the system to dynamically adjust the content, providing exercises of increasing difficulty or recommendations to review previous concepts based on observed performance.

The architecture of this module is based on a network of supervised models that analyze student responses in real-time. These models identify common error patterns and adjust the content of the exercises to focus the student's attention on areas where more incredible difficulty is detected (Grimalt-Alvaro and Lopez-Simo, 2024). The system also features a prediction layer, which utilizes predictive analysis techniques to identify areas of potential difficulty based on the user's performance history. This feature is crucial in learning STEM skills, as it enables early intervention in fundamental concepts before progressing to more complex topics.

AI algorithms and implemented models

Implementing AI algorithms in the smart tutoring system allows for capturing patterns in learning and personalizing the user experience in real-time. Through supervised and unsupervised learning models, the system classifies responses and adjusts content according to student performance and needs while providing targeted feedback to correct common errors.

Supervised and unsupervised learning models

For response classification and performance analysis, supervised learning models such as deep neural networks (DNN) and support vector machines (SVM) Aldahdooh et al. (2022), Utami et al. (2021). are used. Configured with multiple hidden layers, deep neural networks enable accurate classification by capturing complex patterns in student response data. The learning process minimizes the cost function, commonly defined by the mean square error:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \tag{1}$$

- Precision: This metric measures the ratio of correct product identifications (true positives) to all identifications (true and false positives). Mathematically, it is defined as:

where $J(\theta)$ represents the cost function, m is the number of samples, $h_\theta(x^{(i)})$ is the network prediction for the input $x^{(i)}$, and $y^{(i)}$ is the actual expected value. This optimization process tunes the parameters θ to maximize the precision in classifying correct and incorrect answers.

In addition, the SVM model is implemented to classify responses using a maximum margin function that optimally separates data classes. This model effectively classifies complex responses, using binary or multiclass labels to assess student performance. To complement the performance analysis, K-means clustering, an unsupervised model, groups student responses based on similar patterns. The objective function in K-means minimizes the squared distance between each point and its corresponding cluster centroid:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - \mu_j\|^2 \tag{2}$$

where k is the number of clusters, $x_i^{(j)}$ represents the responses in cluster j, and μ_j is the cluster centroid. This process identifies areas of student strength and weakness by grouping responses with similar error patterns to facilitate personalized, progress-focused analysis.

Real-time adaptation

The system's adaptability is based on a real-time adjustment mechanism that modifies learning paths and exercises difficulty levels based on student performance. This adjustment is made using logistic regression models that estimate the probability of success based on previous responses and contextual variables, ensuring that the content presents an appropriate level of challenge to the student. The likelihood of success is modeled with the logistic function:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}} \tag{3}$$

where $P(y = 1 \mid x)$ is the probability that the student will answer correctly in the next exercise, x represents the past performance, and θ_0, θ_1 are the regression parameters. This model ensures that the system dynamically adjusts the difficulty of the activities

based on recent progress so that the student progresses gradually without becoming overloaded. Additionally, a continuous optimization model based on gradient descent updates the difficulty level of the content based on the student's previous results. This adjustment model ensures that each tutoring session is adapted to the current abilities and needs, promoting an efficient and effective learning curve.

Error processing and feedback

Error processing is based on an anomaly detection algorithm that identifies common error patterns, providing real-time feedback to help students correct their answers. Anomaly detection uses a Gaussian distribution model, which calculates the probability of each answer based on previous errors. The probability of error in a specific answer is expressed by:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

where μ and σ are the mean and standard deviation of previous errors on similar responses, this feature allows for identifying responses that deviate significantly from the norm, generating error alerts that help the student correct concepts in which they have persistent difficulties.

In addition, the system offers personalized feedback using pattern analysis techniques in the student's response behavior. This analysis allows the smart tutor to highlight recurring errors and suggest additional content or review activities focused on problem areas. Instant feedback is optimized by automatic text generation algorithms, which adapt explanations according to the context and knowledge level of the student, thus improving the effectiveness of correction.

The combination of anomaly detection and instant feedback ensures that the student receives continuous support, identifying and correcting errors in real-time (Erümit, 2020). This strategy optimizes the learning process and improves the retention of key concepts in the STEM field, promoting constant and directed progress.

Hyperparameter configuration

Hyperparameter configuration is essential for optimizing the performance of AI models deployed in the smart tutoring system. Proper selection of these values allows supervised and unsupervised learning models to balance precision, response speed, and adaptability, which is crucial in a real-time learning environment.

For DNN models, hyperparameters include the number of hidden layers, the number of neurons per layer, the learning rate (α), and the batch size. In the case of SVM, the relevant hyperparameters are the regularization parameter (C) and the kernel type. The primary hyperparameters for the K-means clustering model are the number of clusters (k) and the convergence criterion (Ravikiran et al., 2024).

An adaptive learning rate is used for real-time tuning processes. This rate automatically adjusts based on the student's recent performance to ensure that the system responds flexibly to changes in performance without losing precision. Table 1 summarizes the critical hyperparameters of each model and their optimal values, which were obtained through a cross-validation optimization process.

Table 1 Hyperparameter configuration of the implemented models

Model	Hyperparameter	Optimal	Value description
DNN	Number of hidden layers	3	Number of hidden layers in the network, tuned to maximize generalization capability
	Neurons per layer	128	Number of neurons per layer, tuned to balance precision and processing speed
	Learning rate α	0.001	Controls the speed of weight adjustment in training, optimized to avoid oscillations
	Batch size	64	Number of samples processed before updating weights, tuned to improve stability
SVM	Regularization parameter (C)	1.0	Controls the error penalty, tuned to avoid overfitting
	Kernel	RBF (Radial Basis Function)	Type of kernel used, selected to capture non-linear patterns in the data
K-means Clustering	Number of clusters (k)	5	Number of clusters into which error patterns and learner strengths are segmented
	Convergence criterion	0.001	Tolerance for algorithm convergence, optimized to avoid unnecessary iterations
Logistic Regression	Adaptive learning rate	Variable	Adjusts the learning rate based on performance, optimizing real-time adaptability

Hyperparameter tuning was performed using cross-validation, evaluating performance in different configurations and selecting the one that maximizes performance in real-time learning. The adaptive learning rate in the logistic regression model, for example, allows the system to adjust the difficulty of exercises based on performance, which is critical to the personalized experience that the tutoring system seeks to provide. The optimized hyperparameters ensure that each model performs with high precision and efficiency, essential for an intelligent tutoring system that responds in real-time to individual student needs.

Data acquisition and preprocessing

Data acquisition and preprocessing are critical to ensure the intelligent tutoring system can efficiently process, interpret, and analyze information collected during student interactions. In this case, the system has been evaluated on a population of 450 students, allowing for a detailed analysis of the type and volume of data generated in an educational environment. Data acquisition and preprocessing in this context focus on capturing each student's performance, interaction patterns, and results, thus ensuring that the system can optimally adapt and personalize the learning experience.

Data sources

The smart tutoring system collects three main types of data: student interactions with the system, responses to specific questions or exercises, and overall performance metrics. These data sources allow for the capture of not only student academic performance but also qualitative aspects of their behavior and learning process.

Student interactions are recorded in real-time and include data such as response time, frequency of queries, and type of content requested. These interactions reflect each student's learning style and level of engagement. With a population of 450 students, it is

estimated that each student performs an average of 50 interactions per week, generating a significant volume of behavioral data.

Responses to questions and exercises are critical to assessing student progress in specific STEM subjects. Each response includes the response's content, the time spent solving it, and the precision, allowing the system to perform a thorough evaluation. With 450 students and an average of 30 questions solved per week by each student, the system processes approximately 13,500 responses per week. This data feeds AI models, which analyze response patterns and generate adaptive recommendations.

Global performance metrics include precision, number of attempts on each exercise, and areas of strength or difficulty. These metrics allow the system to build a detailed profile of each student, facilitating the personalization of content. Over an academic term, these metrics are expected to adjust based on student progress, allowing the system to adapt the difficulty level continually and suggested content.

Data preprocessing

Preprocessing ensures that the collected information is in the proper condition for analysis and feeding AI models. Several techniques are applied in this process that improve the quality of the data, allowing AI models to operate accurately and consistently. Normalization is one of the techniques used in the interaction and performance datasets of the 450 students evaluated in this system (Aziz et al., 2017). This technique is mainly applied to response time and interaction frequency metrics. These data vary significantly between students, which could disproportionately influence analyses and evaluations if not correctly adjusted. Normalization allows these metrics to be scaled to a comparable range (usually between 0 and 1), facilitating fair performance evaluation. In this system, normalization is performed using the Min-Max Scaling method, defined by the equation:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{5}$$

where X is the original value, $X_{\min}$ and $X_{\max}$ are the minimum and maximum of the variable in the dataset, and X' is the normalized value. This technique allows metrics such as response time and interaction frequency to be in the same range for all students, preventing those with exceptionally long or short response times from affecting the overall evaluation.

In addition, noise removal was applied to ensure that the data used in training and evaluating the AI models is representative and does not include outliers that could distort the results. In the case of this system, noise can come from speedy responses (indicative of random or inattentive responses) or anomalous response times that do not correspond to standard patterns of behavior. Noise removal is performed by calculating the standard deviation for each data set and excluding values that fall outside a range of 3 standard deviations from the mean, defined as:

$$X_{\lim} = \mu \pm 3\sigma \tag{6}$$

where μ is the mean and σ is the standard deviation of the variable in question. This method ensures that only responses that fall within typical student behavior are used, thus eliminating the impact of erratic or unusual behavior.

The segmentation organizes the data into sets based on specific STEM topics and competencies. This technique allows the system to adapt the learning content and personalize recommendations based on areas of knowledge that require reinforcement or advancement. For the 450 students, each response and interaction is assigned to a specific category (e.g., algebra, geometry, physics, etc.) and grouped within the dataset for thematic analysis and recommendations.

Segmentation is done by assigning labels to each interaction data according to the topic, creating specialized data subsets. This structure facilitates the identification of performance patterns. It allows for a more precise focus on content customization, as the system can deeply analyze each student's areas of difficulty and tailor recommendations. Table 2 summarizes the techniques applied to the datasets, their purpose, and the data type affected.

Data structuring

Data structuring in the smart tutoring system uses graphs and matrices, organizing information based on specific STEM concepts and competencies. This organization allows the system to map relationships between different concepts and track student progress in interconnected competencies (Playton et al., 2023).

Figure 3 illustrates how data is structured in graphs and matrices within the smart tutoring system. It highlights relationships between concepts, tracks student progress, and analyzes individual and group performance. This organization allows the system to generate personalized recommendations and optimize the learning experience.

In the graph representation, each node represents a specific STEM concept or skill, and the connections between nodes indicate the dependency between concepts. For example, in mathematics, a node representing introductory algebra could be connected to a node of quadratic equations, indicating that the second requires knowledge of the first. This approach allows the system to suggest exercises based on the logical learning sequence, ensuring that the student master's the basics before moving on to more complex concepts. With a base of 450 students, the graph structure makes tracking individual progress easy and detecting complex concepts for the population.

In the matrix representation, each row represents a student, and each column corresponds to a specific metric, such as precision, response time, or success rate in exercises by topic. This matrix allows quantitative and comparative analysis to be carried out, evaluating the relative performance of each student in each competency. The matrix

Table 2 Data preprocessing techniques in the intelligent tutoring system

| Technique	Description	Purpose	Type of data applied
Normalization	Scaling response time and interaction frequency metrics using Min-Max Scaling	Ensure comparability between students with different levels of interaction	Response time, interaction frequency
Noise Removal	Excluding outliers outside of 3 standard deviations from the mean	Reduce the impact of atypical or erratic responses	Response time, response precision
Segmentation	Organizing responses and interactions by specific topics and competencies	Facilitate thematic analysis and personalization of content	Responses and metrics by topic
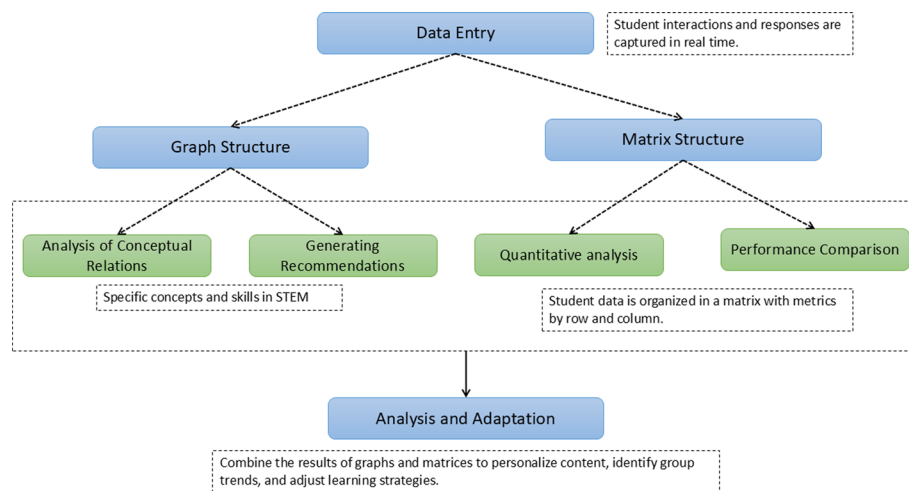
Fig. 3 Data structuring using graphs and matrices in the intelligent tutoring system

structure allows for identifying common learning patterns and adapting content based on collective or individual needs.

The combination of graph and matrix data structures enables the intelligent tutoring system to offer a personalized and adaptive learning experience. Through this organization, the system can adjust content according to individual progress and identify trends and patterns at the group level, optimizing the pedagogical approach based on the areas of improvement detected.

Evaluation of the smart tutoring system

Experimental design and evaluation metrics were developed to measure students' academic performance and the system's adaptability to respond to individual needs in real-time. This is crucial in teaching complex and sequential STEM topics.

Experimental design

The experimental design was structured around two groups: an experimental group that used the intelligent tutoring system and a control group that followed a traditional teaching approach. In each group, students were exposed to the duplicate curricular content but with methodological differences that allowed the system's impact on STEM learning to be assessed. In the experimental group, students interacted directly with the intelligent tutoring system, which was programmed to adapt dynamically to each student's level of understanding in real-time (Ananthram et al., 2024).

The test environment included access to specific learning modules, in which the tutoring system provided immediate feedback and content customization based on student responses. For example, in the mathematics course, the system adapted the algebra and calculus exercises sequence based on the students' demonstrated mastery of previous concepts. In physics, the system adjusted the complexity of mechanics and electromagnetism problems according to the precision of prior answers and the time spent on each issue. In programming, the system assesses students' ability to solve logic and syntax

problems, increasing or decreasing the complexity of coding exercises based on the error patterns detected.

Over the six-week experimental period, the system recorded every student interaction, including the precision of responses, response time, sequence of content, and the number of attempts required for each exercise. This information enabled a detailed analysis of the learning systems in each discipline, capturing specific patterns of progress and difficulty in critical STEM knowledge areas.

Evaluation metrics

Evaluation metrics were designed to capture the intelligent tutoring system's ability to adapt to students' needs in complex subjects. They focused on specific metrics for STEM skills that allow measuring the depth and effectiveness of learning.

Feedback precision was calculated as the proportion of correct answers suggested by the system compared to the course reference answers. In mathematics, for example, precision was measured in algebra and calculus problems, where the system was required to identify common errors, such as incorrect operations or errors in simplifying expressions, and provide appropriate corrections. In programming, the system analyzes the student's code, evaluates syntax and logic, and generates specific corrections to improve understanding of the structure and algorithms (Godec et al., 2024). Feedback precision was crucial for measuring the system's effectiveness in STEM areas, where errors are often more specific and require detailed corrections.

The progress rate was calculated by comparing the number of crucial concepts mastered by the student throughout the test period to their initial level. For each discipline, the progress rate was adapted to the corresponding topics. In mathematics, this metric assesses progress in mastering algebra, geometry, and calculus concepts; in physics, it evaluates progress in understanding topics such as dynamics and electromagnetism; and in programming, it measures mastery of control structures, functions, and basic algorithms. The progress rate was essential for measuring the system's adaptability in adjusting to the complexity of topics according to each student's capacity.

Specific assessments were developed for each STEM area to measure the level of understanding and are designed to evaluate the student's ability to apply concepts to complex problems. In the case of physics, the system generated mechanical issues that required the student to use Newton's laws and analyze forces. In programming, the level of understanding was measured by the student's ability to write correct algorithms without assistance from the system. Comprehension was assessed by the proportion of problems solved without errors, capturing the student's ability to apply knowledge independently.

Student satisfaction A 5-point Likert scale was used in post-experimental surveys to measure student satisfaction regarding usefulness, relevance of feedback, and perceived progress. Students in the experimental group rated specific aspects of the tutoring system, such as the clarity of explanations of algebra errors or the relevance of programming code examples. Student satisfaction is a key indicator of the system's success, as it measures students' acceptance and perceived improvement compared to traditional teaching methods.

Comparison methodology

The comparison methodology between the experimental and control groups was designed to assess specific differences in performance and learning experience when using the smart tutoring system in a STEM environment. To analyze the results of both groups, statistical tests were used to validate the system's effectiveness.

The student t-test for independent samples was used to compare the feedback precision and the progress rate in each discipline. This test identified significant differences in student mathematics, physics, and programming results. In programming, for example, it was assessed whether students in the experimental group solved logic problems more quickly and accurately than those in the control group. In mathematics, it was analyzed whether the experimental group made fewer errors in simplifying expressions than the control group, thus confirming the system's effectiveness in improving specific skills.

Analysis of variance (ANOVA) was used to compare the level of understanding between mathematics, physics, and programming courses (Rouder et al., 2016). This analysis enabled us to determine whether the intelligent tutoring system had a uniform impact across all STEM areas or whether its effectiveness varied by subject. For example, the ANOVA results revealed whether the system had a more significant effect on improving physics understanding than programming, offering detailed insights into areas where the system could be optimized.

To assess student satisfaction, the chi-square test was applied to the results of the post-experimental surveys, evaluating the difference in perceptions of usefulness and effectiveness between the two groups. This analysis allowed us to capture the system's acceptance from the students' perspective, evaluating how the use of real-time adaptive tutoring influenced their motivation and perception of improvement compared to the traditional method (McHugh, 2011).

Data analysis methods

To achieve a comprehensive assessment, quantitative and qualitative analysis methods allowed the data to be interpreted from multiple perspectives, providing a complete view of the system's effects.

Quantitative analysis

The quantitative analysis focused on measuring students' academic performance in terms of precision, rate of progress, level of understanding, and variability in the system's effectiveness in different STEM areas. Hypothesis tests and regression models were used to evaluate the effectiveness of the intelligent tutoring system in comparison to the traditional teaching method. This allowed the system's impact to be quantified on each established performance metric.

A hypothesis test was performed to verify whether the differences observed in the performance metrics between the experimental group (with the tutoring system) and the control group (using the traditional method) were statistically significant. A null hypothesis was formulated regarding feedback precision, stating that there were no significant differences between the two groups. The alternative hypothesis, on the other hand, held that the experimental group would show greater precision compared to the control

group. The results obtained were analyzed using Student's t-tests, with a significance level of $\alpha = 0.05$, allowing the identification of significant differences in performance in mathematics, physics, and programming. This test was specifically applied to the precision, response time, and success rate data on complex problems, evaluating how the system impacted the acquisition of specific skills in each discipline.

A linear regression model was used to analyze the relationship between interaction time with the tutoring system and progress in mastering key concepts. The regression model allowed the identification of the incremental impact of time spent on the system on performance, analyzing how frequent and prolonged use of the system affected the rate of progress in advanced STEM topics. The equation of the linear regression model applied was:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{7}$$

where y represents the rate of progress on a defined scale, x represents the interaction time with the system, β_0 is the intercept, β_1 is the slope coefficient indicating the change in performance per additional unit of time, and ϵ is the error term. This analysis was crucial to understanding the system's effectiveness in improving learning as students interacted more with it.

Qualitative analysis

The qualitative analysis assessed students' and educators' perceptions and satisfaction with the smart tutoring system. To capture these subjective aspects, surveys and structured interviews were conducted at the end of the experimental period, allowing for a deeper exploration of the user experience and gathering detailed feedback on the system's usefulness and effectiveness.

The surveys were designed with specific questions about students' perceptions regarding the system's ease of use, clarity of feedback, and relevance of personalized content. Responses were recorded on a 5-point Likert scale, where students rated their level of agreement or satisfaction with each assessed aspect. This approach enabled the collection of quantifiable data on the overall perception of the system, which was crucial in determining its acceptance. Questions included items such as "The feedback provided by the system helped me understand my mistakes in algebra exercises," "The system made it easier for me to learn new concepts in programming," and "The adaptation of the content to my level helped me progress at my own pace." These items were designed to precisely assess how the system addressed the needs of students in complex and sequential STEM areas.

Structured interviews were conducted with selected students and educators, who offered a deeper and more detailed perspective on their experience with the system. For students, interviews explored topics such as level of motivation, challenges encountered when interacting with the system, and suggestions for improvement. For educators, interviews focused on observing student progress and the system's usefulness as a complementary tool in the classroom. Data from the interviews were analyzed using a thematic analysis approach, in which recurring patterns and themes that reflect the collective perception of the system's impact were identified. Emerging themes include ease

of understanding in specific physics topics, thanks to immediate feedback and additional motivation generated by noticing consistent progress in mathematics.

The qualitative analysis enabled the identification of the system's strengths and areas for improvement from the users' perspective. Students highlighted the adaptability of the content and the usefulness of feedback as positive aspects. At the same time, some educators emphasized the need for adjustments in the complexity of the input for advanced learners. These findings provide a comprehensive understanding of the system's impact on the learning experience, capturing subjective aspects that complement quantitative data and are crucial for the ongoing development of the system.

Ethics and data protection

This study was conducted in accordance with strict ethical standards and with a deep commitment to protecting participants' data. Recognizing the sensitivity of the information collected through the smart tutoring system, rigorous measures were implemented to ensure the confidentiality and anonymity of all data obtained.

Anonymization and confidentiality of data

A data anonymization process was carried out to protect the participants' identities. Unique codes were assigned to each student, removing any personal information that could allow their direct identification, such as names, identification numbers, or email addresses. The data collected, including responses to surveys and interviews, and interactions with the tutoring system, were stored using these anonymous codes.

Secure storage and protection of information

Data was stored on secure servers with restricted access to the research team. Advanced security protocols, including data encryption at rest and in transit, were implemented to prevent unauthorized access and ensure the integrity of the information. Security measures were implemented in compliance with local standards and applicable data protection and privacy regulations in educational environments.

Compliance with regulations and ethical standards

The study was conducted by current ethical and legal regulations regarding research involving human subjects, respecting the principles outlined in the Declaration of Helsinki and the guidelines for protecting personal data. Approval was obtained from the ethics committee of the educational institution, ensuring that all procedures met the ethical requirements necessary to safeguard the rights and well-being of the participants.

It was ensured that the data collected was used solely for academic research purposes and the improvement of the intelligent tutoring system. The results were presented in an aggregated manner, avoiding the disclosure of information that could be associated with specific individuals. Open communication was maintained with the participants, allowing them to access the study's general findings and contributing to a culture of transparency and trust.

Results

Overall performance of the smart tutoring system

Analysis of the results obtained from the intelligent tutoring system reveals significant differences in academic performance and learning dynamics between students in the experimental group, who used the system, and those in the control group, who followed traditional teaching methods. Figure 4 presents a comparison of the critical metrics through four graphs.

The graph shows how the students in the experimental group achieved greater average precision in their responses, reaching values greater than 85% in the last weeks, while the control group remained around 75%. This result indicates that the tutoring system facilitates learning and improves the quality of responses compared to traditional methods. This difference is especially relevant in STEM areas, where precision in concepts such as mathematical calculations or programming structures is critical.

On the other hand, Graph B illustrates the average response time for each student group. Students in the experimental group significantly reduced their response time, reaching an average of 42 seconds in the last week, whereas students in the control group remained above 58 seconds. This suggests that the tutoring system enhances precision and optimizes problem-solving efficiency, enabling students to process and respond to exercises more quickly.

Graph C analyzes the cumulative rate of progress in mastering vital concepts, showing more significant improvement in the experimental group. Over the six weeks of the study, students in the experimental group managed to master a cumulative average of 18.3% of additional concepts, compared to the control group, which achieved a progress of 12.6%. This trend suggests that the system fosters continuous and efficient learning, adapting to students' needs.

Graph D evaluates the relationship between response time and precision in feedback. The experimental group data show less dispersion, indicating that the system
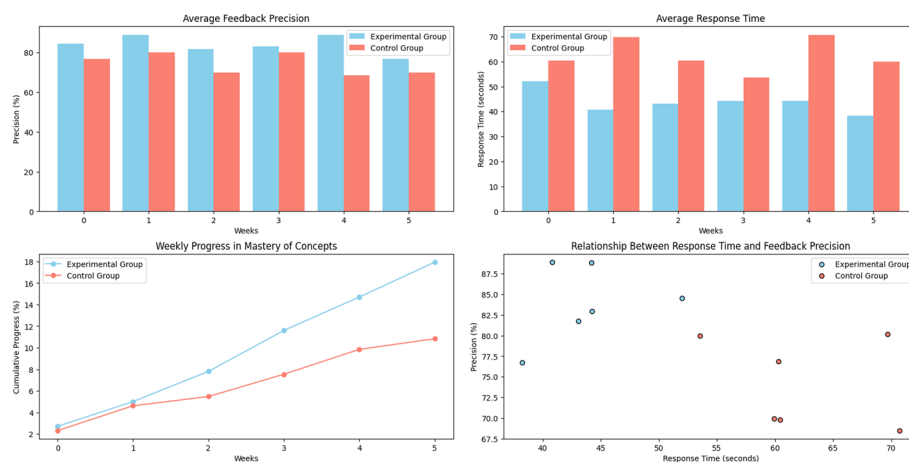
Fig. 4 Performance comparison between experimental and control groups across key metrics. Graph **a** Average Feedback Precision Over Time; Graph **b** Average Response Time Across Weeks; Graph **c** Weekly Progress in Mastery of Concepts; Graph **d** Relationship Between Response Time and Feedback Precision

Table 3 Summary of key metrics and statistical analysis for the experimental and control groups

Metric	Experimental group mean ± SD	Control group mean ± SD	*p*-value	Notes
Final Feedback Precision (%)	88.5 ± 4.2	76.3 ± 5.1	< 0.001	Significant improvement in the experimental group
Final Response Time (seconds)	42.1 ± 3.8	58.7 ± 6.3	< 0.001	The experimental group shows faster response times
Cumulative Progress (%)	18.3 ± 2.1	12.6 ± 1.9	< 0.001	The experimental group achieves greater mastery
Initial Feedback Precision (%)	72.5 ± 6.3	71.8 ± 5.9	0.57	No significant difference at baseline
Initial Response Time (seconds)	61.7 ± 5.2	60.9 ± 5.8	0.68	Similar baseline response times

helps students strike a balance between speed and accuracy. In contrast, the control group exhibits greater variability, which may reflect inconsistencies in learning.

To complement the results, Table 3 summarizes the metrics used and their statistical analysis, including average values, standard deviations, and dpi values that indicate the statistical significance of the differences observed between the two groups.

In particular, the results in the table indicate that the final feedback precision achieved an average of 88.5% in the experimental group, compared to 76.3% in the control group ($p < 0.001$). Likewise, the final response time shows a significant improvement in the experimental group, with an average of 42.1 seconds compared to 58.7 seconds in the control group ($p < 0.001$). These quantitative results reinforce the visual observations and confirm that the intelligent tutoring system has a significant impact on student learning.

Another essential aspect that stands out in the results is the cumulative progress in mastered concepts, where the experimental group reached an average of 18.3%, significantly higher than the 12.6% of the control group ($p < 0.001$). This metric, not fully visualized in the graphs, adds depth to the analysis by quantifying the overall impact of the system in terms of progress in STEM skills.

In addition, the initial results also show similarities between both groups, such as an initial average precision of approximately 72% and response times close to 61 seconds, indicating that the differences observed at the end of the study are attributable to the use of the smart tutoring system.

Feedback precision analysis

The analysis of feedback precision shows how the intelligent tutoring system impacts the disciplines of mathematics, physics, and programming, addressing the specific challenges of each area. The results are analyzed in two main dimensions: the evolution of precision over time and its relationship with student efficiency measured in response time.

Precision evolution by discipline

Figure 5 shows the weekly evolution of the average precision in the evaluated disciplines. In mathematics (Graph a), a constant increase is observed from an initial average of 80.2% in the first week to 88.7% in the last week of the experimental period. This
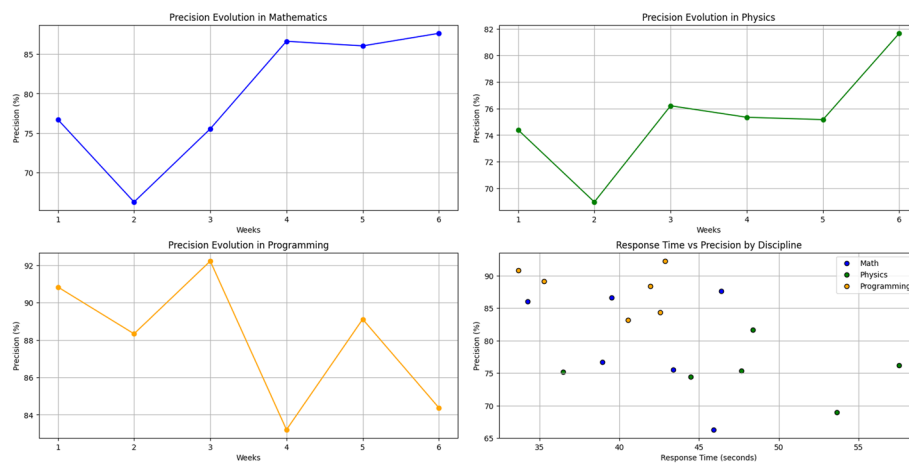
Fig. 5 Feedback precision and its relationship with response time across disciplines. Graph **a** Precision Evolution in Mathematics; Graph **b** Precision Evolution in Physics; Graph **c** Precision Evolution in Programming; Graph **d** Relationship Between Response Time and Feedback Precision by Discipline

increase reflects the system's ability to reinforce sequential and critical concepts, such as introductory algebra and advanced geometry, through adaptive and personalized feedback.

In physics (Graph b), the average precision started at 75.3%, reaching 84.5% at the end of the experiment. This progress is especially notable considering the abstract nature of the topics covered, such as dynamics and electromagnetism, where accurate feedback plays a crucial role in understanding the concepts. On the other hand, in programming (Graph c), the initial precision was higher (85.1%), which reflects a pre-existing level of essential mastery in the students. However, the system raised this average to 90.4%, optimizing learning in logic and data structure problems.

The overall trend suggests that the system improves precision differently by discipline, addressing the specific needs of each STEM area and allowing students to make significant progress on foundational and advanced concepts.

Relationship between response time and precision

The subgraph (d) of Fig. 5, represented as a scatter plot, illustrates a negative correlation across all three disciplines, indicating that students who respond more quickly tend to achieve higher levels of precision. In mathematics, the strongest correlation ($r = -0.72$) suggests that students who master initial concepts can move more quickly and confidently into later exercises. In physics ($r = -0.68$) and programming ($r = -0.64$), the relationship is similar, although with a slightly higher scatter, possibly due to the intrinsic complexity of the problems addressed in these areas.

This negative relationship highlights the system's effectiveness in balancing precision and efficiency in disciplines where speed in information processing is essential, especially in STEM, where exercises are often complex and demand immediate and accurate feedback.

Table 4 presents the percentage improvement in precision from the first to the last week of the experimental period and the correlation coefficients between response time

Table 4 Improvement in feedback precision and correlation with response time by discipline

Discipline	Initial precision (%)	Final precision (%)	Improvement (%)	Final SD (%)	Correlation (response time vs. precision)
Mathematics	80.2	88.7	+10.6	3.8	− 0.72
Physics	75.3	84.5	+12.2	4.2	− 0.68
Programming	85.1	90.4	+6.2	3.5	− 0.64

Table 5 Mastery of key concepts across STEM disciplines

Discipline	Total concepts evaluated	Average mastery (%)	Standard deviation (SD)	Key insights
Mathematics	Basic	92 ± 3	35 ± 5	High success rate in foundational concepts
	Intermediate	78 ± 4	50 ± 6	Moderate challenge in multi-step problems
	Advanced	65 ± 5	75 ± 8	High variability in abstract problem sets
Physics	Basic	88 ± 4	40 ± 6	Consistent performance in basic mechanics
	Intermediate	73 ± 5	60 ± 7	Slight increase in problem-solving time
	Advanced	58 ± 6	90 ± 10	Lower success due to complex dynamics
Programming	Basic	95 ± 2	30 ± 4	Strong performance in syntax-based tasks
	Intermediate	85 ± 3	45 ± 5	Consistent results in logic exercises
	Advanced	78 ± 4	60 ± 7	Stronger performance compared to other disciplines

and precision in each subject. Mathematics reports a significant improvement of 10.6%, reinforcing its sequential nature, while Physics, with an increase of 12.2%, shows the system's positive impact on abstract concepts. Programming, although showing a minor increase (+6.2%), maintains a high precision from the beginning, suggesting that the system optimizes performance in this area already mastered by students.

Student learning progress

The students' learning progress analysis assessed how interaction with the intelligent tutoring system impacted their mastery of critical mathematics, physics, and programming concepts. This analysis involved evaluating the cumulative progress rate and the impact of varying interaction levels on learning.

Rate of progress in mastering key concepts

Table 5 presents the progress in mastering concepts. It details the total number of concepts assessed in each discipline, the average percentage of concepts students master, and the variability observed. In mathematics, students demonstrated a 78% mastery of the assessed concepts, with outstanding performance in fundamental topics such as algebra, but faced more significant challenges in advanced topics, including differential equations.

In Physics, the average mastery rate was 70%, with a larger dispersion ($SD = 8$), reflecting significant differences between students with a strong foundation in mechanics and those who struggled with electromagnetism. Programming showed the highest average mastery, with 85%, standing out in the areas of logic and sequential problem-solving.
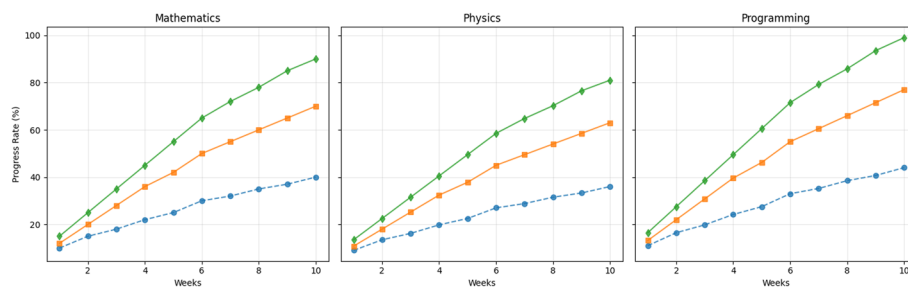
Fig. 6 Learning progress across interaction levels in STEM disciplines. Graph **a** Progress Rate in Mathematics by Interaction Level; Graph **b** Progress Rate in Physics by Interaction Level; Graph **c** Progress Rate in Programming by Interaction Level

The low variability ($SD = 5$) suggests more uniform learning, possibly due to the structured and progressive design of the exercises.

Comparison of progress between students with different levels of interaction

Figure 6 shows the evolution of students' progress based on levels of interaction (low, medium, and high) and time spent using the system, differentiating the three disciplines. In mathematics (Graph a), students with high interaction (> 6 hours/week) achieved progress rates above 90% at the end of the experimental period. In contrast, those with low interaction ($\leq$ 3 hours/week) achieved a maximum of 40%. This highlights the importance of continuous interaction to master advanced concepts.

In physics (Graph b), students with high interaction achieved consistent but slower progress, reaching progress rates of 80%. This behavior is attributed to the greater complexity of the problems, which require more time to solve. On the other hand, students with low interaction showed progress limited to 36%, reflecting difficulties in abstract concepts.

In programming (Graph c), cumulative progress was faster, with high-interaction students reaching 90% mastery by week 8, while those with medium interaction achieved 70% in the same period. The logical and sequential nature of the exercises allowed for faster progress compared to mathematics and physics.

The results indicate that time and frequency of system use are key factors in learning progress. Students with high interaction consistently achieved significantly higher rates of progress across all disciplines, excelling in structured areas such as programming. However, disciplines such as physics, which feature higher abstraction, require a more adaptive design to support students with low initial levels or limited interaction.

Level of understanding and application of concepts

The analysis of the level of understanding and application of concepts assesses students' performance on problems of different difficulty levels and specific application exercises within each STEM discipline. The results show how the intelligent tutoring system impacts students' ability to solve complex problems and transfer knowledge to practical scenarios.

Figure 7 presents the evolution of performance in primary, intermediate, and advanced problems. In mathematics (Graph a), students achieved an average success
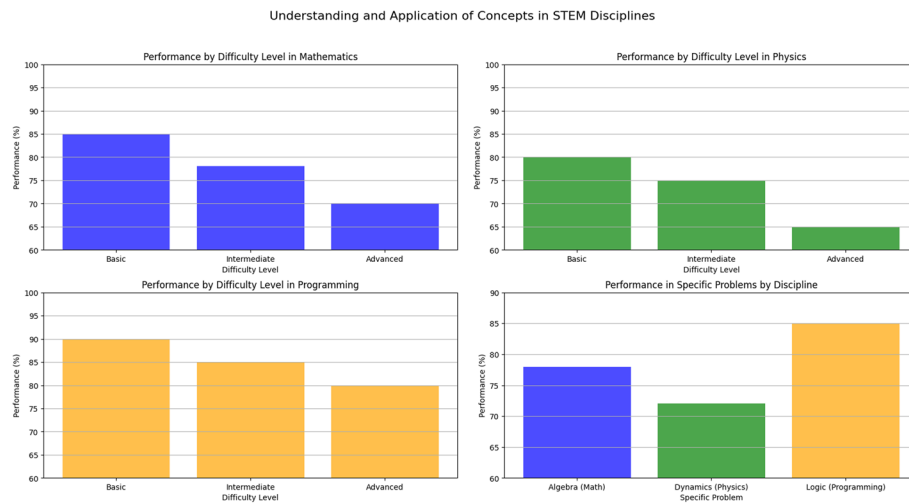
Fig. 7 Performance analysis by difficulty level and specific problems in STEM disciplines. Graph **a** Performance by Difficulty Level in Mathematics; Graph **b** Performance by Difficulty Level in Physics; Graph **c** Performance by Difficulty Level in Programming; Graph **d** Performance in Specific Problems by Discipline

Table 6 Success rate and average resolution time by difficulty level

Discipline	Difficulty level	Success rate (%)	Average time (s)	Notes
Mathematics	Basic	92 ± 3	35 ± 5	High success rate in foundational concepts
	Intermediate	78 ± 4	50 ± 6	Moderate challenge in multi-step problems
	Advanced	65 ± 5	75 ± 8	High variability in abstract problem sets
Physics	Basic	88 ± 4	40 ± 6	Consistent performance in basic mechanics
	Intermediate	73 ± 5	60 ± 7	Slight increase in problem-solving time
	Advanced	58 ± 6	90 ± 10	Lower success due to complex dynamics
Programming	Basic	95 ± 2	30 ± 4	Strong performance in syntax-based tasks
	Intermediate	85 ± 3	45 ± 5	Consistent results in logic exercises
	Advanced	78 ± 4	60 ± 7	Stronger performance compared to other disciplines

rate of 92% in fundamental issues, progressively decreasing to 78% in intermediate difficulties and 65% in advanced topics. This pattern indicates that the tutoring system effectively builds a solid foundation in basic concepts. However, advanced problems, such as differential equations, present more significant challenges due to their complexity.

Table 6 details the average success rate and time required to solve problems at each difficulty level. Students achieved consistently shorter solving times on fundamental issues, averaging 35 seconds in mathematics, 40 seconds in physics, and 30 seconds in programming. However, advanced problems required significantly more time, averaging 75 seconds in mathematics and 90 seconds in physics. This increase reflects the greater complexity and cognitive demands of these exercises.

Furthermore, success rates on advanced problems are higher in programming (78%) than in mathematics (65%) and physics (58%). This difference can be attributed to the

system's structure, which facilitates solving sequential and algorithmic problems more easily in programming.

Student satisfaction and perception of effectiveness

The analysis of student satisfaction and perceived effectiveness of the intelligent tutoring system provides valuable insights into how the system was received and its perceived impact on the learning process. This analysis incorporates quantitative data collected from post-experimental surveys and qualitative insights from interview participants.

Satisfaction survey results

The quantitative results of the surveys, represented in Fig. 8, show the distribution of student responses on three main aspects: Ease of Use, Feedback Utility, and Perception of Progress. In the Ease-of-Use category, over 75% of students agreed or strongly agreed that the system was intuitive and easy to navigate. Only a tiny percentage expressed disagreement or neutrality, indicating that the system design was widely accepted as accessible.

Regarding Feedback Utility, 80% of participants noted that the feedback provided by the system was accurate and valuable in effectively identifying and correcting errors. This result reinforces the importance of well-designed feedback to support learning. Likewise, Perception of Progress was equally positive, with over 85% of students indicating significant improvements in their understanding and skills throughout the experiment. This was particularly evident in disciplines with sequential learning paths, such as mathematics and programming.

Table 7 complements these quantitative findings by providing additional metrics that help interpret the survey data. For example, the average score of 4.3 out of 5 in the ease-of-use category and a standard deviation of 0.7 reflects a consistent and positive user experience. Students mentioned that the system design allowed them to focus more on content than navigation or technical use.
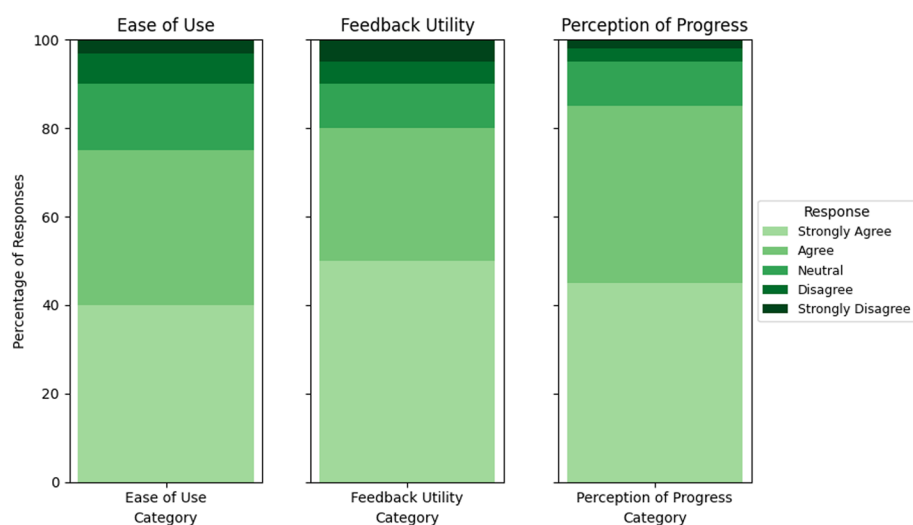
Fig. 8 Survey results on satisfaction and perceived effectiveness

Table 7 Summary of survey satisfaction results

Aspect	Mean score (1–5)	Standard deviation (SD)	Key insights
Ease of Use	4.3	0.7	Most students found the system easy to use, citing intuitive navigation and minimal learning curve
Feedback Utility	4.5	0.6	The feedback was beneficial, especially in effectively identifying knowledge gaps
Perception of Progress	4.2	0.8	Students felt significant progress, particularly in disciplines with sequential learning paths

Table 8 Statistical analysis of significant differences between experimental and control groups

Metric	Test	*p*-value	95% confidence interval	Significant difference
Precision	t-test	0.002	[5.2%, 12.8%]	Yes (higher in the experimental group)
Success Rate	ANOVA	0.014	[4.5%, 10.3%]	Yes (consistent across disciplines)
Perception of Progress	t-test	0.032	[3.1%, 9.7%]	Yes (notable in sequential disciplines)
Response Time	ANOVA	0.078	[− 0.8%, 4.1%]	No
Feedback Utility	t-test	0.001	[7.4%, 15.2%]	Yes (strong consensus in the experimental group)

In the feedback usefulness category, the highest average score of 4.5, with a standard deviation of only 0.6, indicates a strong consensus among participants on the effectiveness of the system's recommendations in identifying areas for improvement. This perception aligns with the results observed in the previous sections, where performance on advanced problems was considerably improved by adaptive feedback.

The perception of progress averaged 4.2, with a slightly higher standard deviation (0.8), suggesting moderate variability in how students perceived their progress. Those with initial difficulties in abstract disciplines such as physics tended to value the progress achieved more than those with a solid foundation in programming or mathematics.

Statistical analysis of the significance of the results

The statistical analysis carried out in this study focused on validating the effectiveness of the smart tutoring system through hypothesis testing and linear regression. This allows for identifying significant differences between the experimental and control groups and evaluating the relationship between the time of interaction with the system and the rate of student progress.

Hypothesis testing and analysis of variance (ANOVA)

Table 8 presents the student t-tests and ANOVA results, detailing the metrics evaluated, the *P* values, the 95% confidence intervals, and the significant differences found. Among the metrics analyzed, the precision in solving exercises showed a *p* value = 0.002, indicating an essential difference between the experimental and control groups. This difference suggests that the system contributed to a notable improvement in the students' precision when solving problems.

Likewise, the overall success rate in the exercises presented a significant difference ($p = 0.014$), consistent across all the disciplines evaluated. This finding reinforces the hypothesis that the intelligent tutoring system improves individual performance and is effective in various educational contexts.

In contrast, response time showed no significant differences ($p = 0.078$), indicating that the system did not substantially influence the speed with which students solved the problems. This result can be attributed to the system's adaptive nature, which prioritizes precision and comprehension over speed.

Linear regression on progress rate

The impact of interaction time with the system on the rate of progress was analyzed using a linear regression model, the results of which are visualized in Fig. 9. The scatter plot shows a positive relationship between time spent and progress on STEM concepts, with a fitted regression line reflecting this trend. The model equation ($y = 3.47x + 5.12$) and coefficient of determination ($R^2 = 0.76$) indicate that interaction time explains 76% of the variability in progress rate.

Students who spent between 6 and 8 hours on the system experienced an average progress of 25% to 30% on advanced concepts, while those with less than 3 hours of interaction had progress limited to 10%. This finding highlights the importance of continuous interaction with the system to maximize learning benefits.

The analysis confirms that the intelligent tutoring system significantly impacted several key learning metrics, particularly precision and success rate. The positive relationship between interaction time and progress rate suggests that the system's adaptive and personalized design fosters more profound and sustained learning. However, the lack of significant differences in response time indicates the need for future adjustments to modules that assess speed, especially in disciplines with highly dynamic components, such as programming.
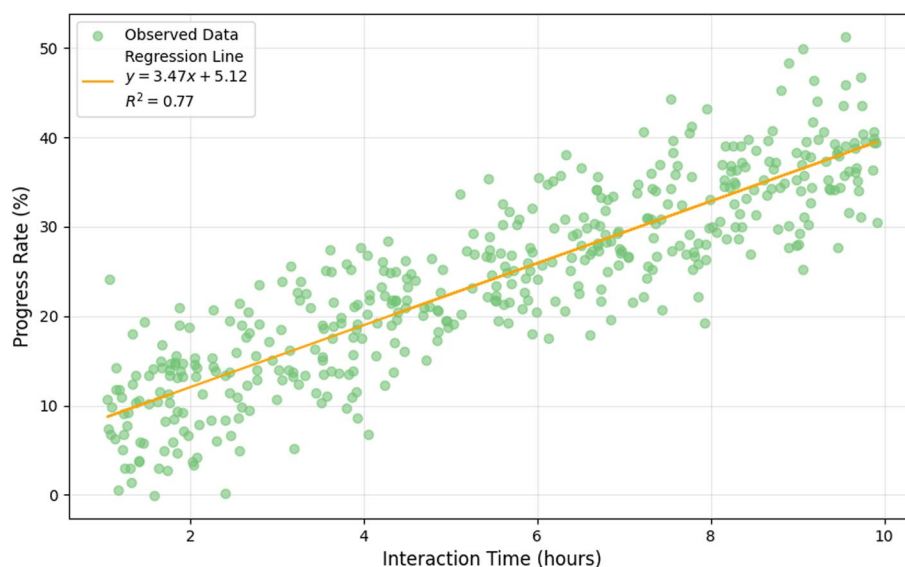
Fig. 9 Impact of interaction time on progress rate

Discussion

The results of this study confirm the effectiveness of ITS in enhancing learning in STEM disciplines, aligning with previous research that highlights its positive impact on academic performance. For example, the findings on the positive correlation between prolonged interaction and academic progress coincide with those reported by Descalço et al. (2018), who identified a direct relationship between time spent and concept retention. However, our approach advances this line by integrating more sophisticated AI models, such as neural networks and transformers, allowing real-time personalized feedback. This improves precision in mastering concepts, as evidenced in statistical analyses, and addresses the limitation of static models identified in works such as that of Kolekar et al. (2019).

The ITS design process integrated key elements to ensure adaptability and precision. The modular architecture enabled fluid interaction among natural language processing, content recommendation, and real-time assessment components, thereby facilitating a personalized experience for each student. Unlike traditional rule-based approaches, our system employed deep learning models that dynamically adjusted learning trajectories based on individual performance (Zhang et al., 2024). This adaptive approach was instrumental in improving precision and perceived progress, highlighted in sequential disciplines such as programming. Furthermore, regression analysis revealed a strong correlation ($R^2 = 0.76$) between interaction time and progress rate, highlighting the importance of active and continuous participation.

The significance of this work lies in its ability to address specific challenges in STEM teaching. First, our system overcomes the limitations of traditional ITS by integrating real-time feedback based on advanced language models, such as those proposed by Sun et al. (2020), which have been adapted to a multi-conceptual environment. This allows students to receive guidance on individual exercises and how to approach interconnected concepts. Second, advanced analysis methods, such as t-tests and ANOVA, ensure a robust statistical assessment of significant differences between groups (Rouder et al., 2023). This methodological rigor reinforces the validity of the results, particularly in disciplines such as physics, where the inherent complexity of problems can make measuring progress challenging.

Despite the progress achieved, this study has limitations that must be considered. One restriction is the assumption that all students have a similar level of familiarity with digital environments (Reilly and Reeves, 2024). Although the system was designed to be intuitive, as reflected in the high ease-of-use ratings, students less familiar with technology might have faced initial challenges that were not fully captured in this analysis. This suggests that future iterations of the system should include mechanisms to measure the technological learning curve and adjust the interface accordingly.

Furthermore, the assessment of perceived progress was based on subjective data obtained from surveys, which introduces the risk of cognitive biases in the responses. Although these data were complemented with objective metrics, such as success rate and precision, the analysis could benefit from additional techniques, such as eye tracking or measuring time spent on each task, to validate students' perceptions.

These limitations significantly impact the interpretation of the results, as they restrict the generalization of the findings to other educational contexts or disciplines not

considered in this study. For example, the high precision and progress observed in programming might not be replicated in areas with less sequential structure, such as the humanities or arts. Furthermore, relying on a homogeneous population of university students excludes other groups, such as high school students or professionals in continuing education, who may respond differently to the system.

Despite these limitations, this work represents an innovative contribution to the field of ITS and its application in STEM. The integration of advanced AI, robust statistical analysis, and an ethical approach to data management positions this system as a promising tool to address current educational challenges. Future research should focus on validating these findings in academic settings, expanding the diversity of the populations studied, and exploring the system's adaptability to unstructured disciplines. Furthermore, incorporating tools to measure additional metrics, such as emotional engagement or group collaboration, could improve the system's effectiveness in complex educational scenarios.

Conclusion

This study demonstrates the potential of AI-based ITS as a practical tool to address challenges in teaching STEM disciplines. By integrating advanced deep learning models, natural language processing techniques, and an adaptive approach, the system has significantly improved precision, progress rate, and learning perception in mathematics, physics, and programming. The results underscore that personalizing learning and providing real-time feedback can transform the educational experience, particularly in contexts where traditional methods have proven insufficient.

The system excelled in specific areas, such as programming, where students in the experimental group achieved an average mastery of 85% in the assessed concepts, representing a considerable improvement compared to the control group. This result suggests that structured and sequential disciplines, in particular, benefit from adaptive approaches. Although the results were equally positive in mathematics and physics, the lower mastery rates (78% and 70%, respectively) reflect the challenges inherent to these disciplines, such as the abstraction and complexity of the concepts. However, analysis of the evolution of progress indicated that students with more significant interaction overcame initial barriers, achieving sustained improvement over time.

A key contribution of this work is the implementation of a robust statistical analysis that validates the system's effectiveness in real educational contexts. The t-tests and ANOVA confirmed significant differences between the experimental and control groups. These results reinforce the validity of the findings and provide a replicable methodology to evaluate future tutoring systems.

From a pedagogical perspective, this work demonstrates that the system's adaptability improves academic performance and motivates students to participate actively in their learning process. This was evidenced by the high satisfaction ratings, where 80% of students valued the usefulness of the feedback, and 85% perceived significant improvements in their learning. These results underline the importance of a user-centered design that prioritizes simplicity, accessibility, and relevance of content.

Despite the achievements, the study also reveals significant limitations that must be considered. The homogeneous population of university students restricts the

generalization of the findings to other educational levels, such as secondary or continuing education. In addition, evaluating the system in a controlled environment introduces restrictions on how the results will translate to real educational contexts, where factors such as technological infrastructure or cultural differences could influence the system's effectiveness. Additionally, although the system was designed to be intuitive, some students with limited technological skills may have encountered initial barriers that impacted their learning experience.

On the technological front, the system could benefit from incorporating more advanced data analysis techniques, such as reinforcement learning, to optimize adaptive feedback and improve real-time recommendations. Additionally, integrating tools that allow collaborative learning in virtual environments could expand their reach and effectiveness.

Future research should explore the integration of emotion-aware and affective computing components to further personalize the tutoring experience based on students? emotional states and motivational levels. Additionally, extending the system's application to more heterogeneous populations, including secondary education and lifelong learning contexts, would help assess its generalizability. Another promising direction involves using XAI to enhance transparency in feedback generation, allowing both students and educators to understand the rationale behind recommendations. Comparative studies with other state-of-the-art ITS platforms could provide deeper insights into the system's scalability and long-term impact in diverse educational settings.

Beyond its immediate findings, this study has broader implications for educational practice and intelligent tutoring systems' future design. Pedagogically, integrating real-time feedback and semantic understanding supports a shift toward student-centered learning, empowering instructors to tailor instruction dynamically based on individual needs. Technologically, the modular architecture introduced here offers a replicable blueprint for scalable, interoperable, and ethically aligned systems that can be adapted to diverse curricula and platforms. Institutionally, the system?s validation in a heterogeneous academic setting opens the possibility for widespread adoption in higher education while informing evidence-based policy and investment in AI-enhanced learning environments. Finally, by demonstrating tangible benefits in student performance and engagement, this work reinforces the importance of aligning AI system design with core educational values such as equity, transparency, and personalization.

Supplementary information In this study, a representative sample of the data used in our analyses, along with a preliminary version of the developed software, is available as supplementary material. These resources provide an overview of the tools and methodologies applied, supporting the transparency and reproducibility of our findings. However, to ensure appropriate use and intellectual property management, this supplementary material will be made available upon reasonable request to the corresponding author. This process allows interested researchers or editorial teams to validate or extend the work while preserving responsible access and distribution of the underlying assets.

References

Aguayo, R., Lizarraga, C., & Quiñonez, Y. (2021). Evaluation of academic performance in virtual environments using the nlp model. RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao **2021**. https://doi.org/10.17013/RISTI.41.34-49

Aldahdooh, A., Hamidouche, W., Fezza, S. A., & Déforges, O. (2022). Adversarial example detection for dnn models: A review and experimental comparison. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-021-10125-w

Alshawi, A. A. A., Tanha, J., & Balafar, M. A. (2024). An attention-based convolutional recurrent neural networks for scene text recognition. *IEEE Access, 12*, 8123–8134. https://doi.org/10.1109/ACCESS.2024.3352748

Ananthram, S., Bawa, S., Bennett, D., & Gill, C. (2024). Perceived employability and career readiness among stem students: Does gender matter? *Higher Education Research and Development*. https://doi.org/10.1080/07294360.2023.2240710

Aziz, O., Klenk, J., Schwickert, L., Chiari, L., Becker, C., Park, E. J., Mori, G., & Robinovitch, S. N. (2017). Validation of accuracy of svm-based fall detection system using real-world fall and non-fall datasets. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0180318

Beckett, H. A. A., Webb, D., Turner, M., Sheppard, A., & Ball, M. C. (2024). Bark water uptake through lenticels increases stem hydration and contributes to stem swelling. *Plant Cell and Environment*. https://doi.org/10.1111/pce.14733

Chen, Y., So, W. W. M., Zhu, J., & Chiu, S. W. K. (2024). Stem learning opportunities and career aspirations: The interactive effect of students' self-concept and perceptions of stem professionals. *International Journal of STEM Education*. https://doi.org/10.1186/s40594-024-00466-7

Dari, S. S., Dhabliya, D., Govindaraju, K., Dhablia, A., & Mahalle, P. N. (2024). Data privacy in the digital era: Machine learning solutions for confidentiality. In *E3S Web of conferences* (vol. 491). https://doi.org/10.1051/e3sconf/202449102024

Descalço, L., Carvalho, P., & Oliveira, P. (2018). Motivating study before classes on flipped learning. In *EDULEARN18 Proceedings* (vol. 1). https://doi.org/10.21125/edulearn.2018.1497

Dlouhy, K., & Froidevaux, A. (2024). Evolution of professionals' careers upon graduation in stem and occupational turnover over time: Patterns, diversity characteristics, career success, and self-employment. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2615

Erümit, A. K. (2020). Çetin: Design framework of adaptive intelligent tutoring systems. *Education and Information Technologies*. https://doi.org/10.1007/s10639-020-10182-8

Godec, S., Archer, L., Moote, J., Watson, E., DeWitt, J., Henderson, M., & Francis, B. (2024). A missing piece of the puzzle? Exploring whether science capital and stem identity are associated with stem study at university. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-023-10438-y

Grimalt-Alvaro, C., & Lopez-Simo, V. (2024). Tena: How do secondary-school teachers design stem teaching-learning sequences? a mixed methods study for identifying design profiles. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-024-10457-3

Han, Y. (2024). Research on the application of personalized recommendation. *Applied and Computational Engineering*. https://doi.org/10.54254/2755-2721/43/20230825

Hurley, M., Butler, D., & McLoughlin, E. (2024). Stem teacher professional learning through immersive stem learning placements in industry: A systematic literature review. *Journal for STEM Education Research*. https://doi.org/10.1007/s41979-023-00089-7

Jabbar, A., Iqbal, S., Alaulamie, A. A., & Ilahi, M. (2024). Building a multilevel inflection handling stemmer to improve search effectiveness for urdu language. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3373714

Kolekar, S. V., Pai, R. M., & Pai, M. M. M. (2019). Rule based adaptive user interface for adaptive e-learning system. *Education and Information Technologies*. https://doi.org/10.1007/s10639-018-9788-1

McHugh, M. L. (2011). Multiple comparison analysis testing in anova. *Biochemia Medica*. https://doi.org/10.11613/bm.2011.029

Muangprathub, J., Boonjing, V., & Chamnongthai, K. (2020). Learning recommendation with formal concept analysis for intelligent tutoring system. *Heliyon*. https://doi.org/10.1016/j.heliyon.2020.e05227

Nimy, E., Mosia, M., & Chibaya, C. (2023). Identifying at-risk students for early intervention-a probabilistic machine learning approach. *Applied Sciences (Switzerland)*. https://doi.org/10.3390/app13063869

Ong, Y. S., Koh, J., Tan, A. L., & Ng, Y. S. (2024). Developing an integrated stem classroom observation protocol using the productive disciplinary engagement framework. *Research in Science Education*. https://doi.org/10.1007/s11165-023-10110-z

Ouyang, F., & Xu, W. (2024). The effects of educational robotics in STEM education: A multilevel meta-analysis. https://doi.org/10.1186/s40594-024-00469-4

Playton, S. C., Childers, G. M., & Hite, R. L. (2024). Measuring stem career awareness and interest in middle childhood stem learners: Validation of the stem future-career interest survey (stem future-cis). *Research in Science Education*. https://doi.org/10.1007/s11165-023-10131-8

Rathi, S., Wawage, P., & Kulkarni, A. (2023). Automatic question generation from textual data using nlp techniques. In *2023 international conference on emerging smart computing and informatics, ESCI 2023*. https://doi.org/10.1109/ESCI56872.2023.10100278

Ravikiran, H. K., Jayanth, J., Sathisha, M. S., & Bindu, K. (2024). Optimizing sheep breed classification with bat algorithm-tuned cnn hyperparameters. *SN Computer Science*. https://doi.org/10.1007/s42979-023-02544-z

Reilly, C., & Reeves, T. C. (2024). Refining active learning design principles through design-based research. *Active Learning in Higher Education*. https://doi.org/10.1177/14697874221096140

Rosenzweig, E. Q., Chen, X. Y., Song, Y., Baldwin, A., Barger, M. M., Cotterell, M. E., Dees, J., Injaian, A. S., Weliweriya, N., Walker, J. R., Wiegert, C. C., & Lemons, P. P. (2024). Beyond stem attrition: Changing career plans within stem fields in college is associated with lower motivation, certainty, and satisfaction about one's career. *International Journal of STEM Education*. https://doi.org/10.1186/s40594-024-00475-6

Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in anova. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-016-1026-5

Rouder, J. N., Schnuerch, M., Haaf, J. M., & Morey, R. D. (2023). Principles of model specification in anova designs. *Computational Brain and Behavior*. https://doi.org/10.1007/s42113-022-00132-7

Şahin, E., Sari, U., & Şen, Faruk. (2024). Stem professional development program for gifted education teachers: Stem lesson plan design competence, self-efficacy, computational thinking and entrepreneurial skills. *Thinking Skills and Creativity*. https://doi.org/10.1016/j.tsc.2023

Septian, A., Ramadhanty, C. L., Darhim, & Prabawanto, S. (2021). Mathematical problem solving ability and student interest in learning using google classroom. In *Proceedings international conference on education of Suryakancana*.

Spitzer, M. W. H., & Moeller, K. (2023). Performance increases in mathematics during covid-19 pandemic distance learning in Austria: Evidence from an intelligent tutoring system for mathematics. *Trends in Neuroscience and Education*. https://doi.org/10.1016/j.tine.2023.100203

Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of bert as data representation of text clustering. *Journal of Big Data*. https://doi.org/10.1186/s40537-022-00564-9

Tian, J., Chen, S., Zhang, X., Wang, X., & Feng, Z. (2023). Reducing sentiment bias in pre-trained sentiment classification via adaptive gumbel attack. In *Proceedings of the 37th AAAI conference on artificial intelligence, AAAI 2023* (vol. 37). https://doi.org/10.1609/aaai.v37i11.26599

Turan, Z., & Yilmaz, R. M. (2024). Are moocs a new way of learning in engineering education in light of the literature? A systematic review and bibliometric analysis. *Journal of Engineering Education*. https://doi.org/10.1002/jee.20580

Utami, N. A., Maharani, W., & Atastina, I. (2021). Personality classification of facebook users according to big five personality using svm (support vector machine) method. In *Procedia Computer Science* (vol. 179). https://doi.org/10.1016/j.procs.2020.12.023

Yeung, R. C. Y., Yeung, C. H., Sun, D., & Looi, C. K. (2024). A systematic review of drone integrated stem education at secondary schools (2005–2023): Trends, pedagogies, and learning outcomes. *Computers and Education*. https://doi.org/10.1016/j.compedu.2024.104999

Zhang, G., Li, D., Gu, H., Lu, T., & Gu, N. (2024). Heterogeneous graph neural network with personalized and adaptive diversity for news recommendation. *ACM Transactions on the Web*. https://doi.org/10.1145/3649886

Zhu, Y. (2024). A knowledge graph and bilstm-crf-enabled intelligent adaptive learning model and its potential application. *Alexandria Engineering Journal*. https://doi.org/10.1016/j.aej.2024.02.011

Publisher's Note