

Mutation-Based Error Diagnosis in a Web Intelligent Tutoring System for First-Degree Equations: A Quasi-Experimental Study

Alan de Oliveira Santana¹, Thiago Reis da Silva², Eduardo Henrique da Silva Aranha¹

¹Departamento de Informática e Matemática Aplicada (DIMAp)
Universidade Federal do Rio Grande do Norte – UFRN

²Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA)
alandeoliveirasantana@gmail.com, thiago.reis@ifma.edu.br, eduardoaranha@dimap.ufrn.br

Abstract. *This paper investigates the use of mutation analysis as a mechanism for diagnosing student errors in a Web-based Intelligent Tutoring System for first-degree equations. The proposed module generates alternative solution trajectories by injecting atomic error patterns into correct solution steps and comparing the resulting answers with student responses. A quasi-experimental study involving 113 participants was conducted using a simulated test with ten procedurally generated problems. The mutation engine identified 70 of 387 incorrect answers (18.09%), while the guided step-by-step interaction achieved a 75.94% accuracy rate (101 correct responses out of 133). Results indicate that mutation-based reasoning reconstruction can provide interpretable diagnostic feedback and support guided problem solving. The paper discusses the potential, limitations, and future applications of mutation-based diagnosis in Intelligent Tutoring Systems.*

1. Introduction

The use of computing in education has been extensively investigated due to its potential to support engagement, personalization, and learning outcomes [Smith et al. 2023, Zerkouk et al. 2025]. Among the technologies proposed for this purpose, Intelligent Tutoring Systems (ITS) have been applied in domains such as distance education, language learning, and mathematics education, supporting problem solving, learner monitoring, and adaptive feedback [Alam 2023, Crossley et al. 2017, Kautzmann and Jaques 2019, Ruan et al. 2020].

A central objective of ITS is to identify student misconceptions and provide personalized instructional support [Gong and Beck 2015, Xing and Goggins 2015, Lin et al. 2023]. Although Bayesian student models, learning analytics, worked examples, and step-by-step tutoring have shown promising results [Gamboa and Fred 2002], accurately associating an incorrect answer with the reasoning process that produced it remains an open challenge in mathematical problem solving.

Mutation analysis, originally proposed for software testing [Untch et al. 1993], has subsequently been applied to domains such as deep neural network testing [Wang et al. 2021] and compiler validation [Shen et al. 2021]. Inspired by this principle, this work investigates mutation analysis as a mechanism for reconstructing plausible student reasoning paths in a Web-based ITS for first-degree equations. Mutation operators

derived from recurring mathematical misconceptions are injected into intermediate solution steps to generate alternative reasoning trajectories, which are compared with student responses. When a generated trajectory reproduces the student's answer, the system associates it with candidate misconceptions and provides interpretable diagnostic feedback, preserving the transparency of explicit reasoning reconstruction while enabling future integration with Artificial Intelligence techniques [de Sousa and Cruz 2024].

To evaluate the feasibility of the proposed approach, a mutation-based module was integrated into a Web ITS and assessed through a quasi-experimental study involving 113 participants. The main contributions of this work are: (i) a mutation-based mechanism for reconstructing plausible student reasoning paths from atomic error patterns; (ii) its integration into a Web-based mathematics tutoring environment; and (iii) an empirical evaluation discussing the potential and limitations of mutation-based reasoning reconstruction in educational settings.

To guide this investigation, the following research questions were defined:

- **RQ1:** To what extent can mutation-based reasoning reconstruction associate student answers with plausible error trajectories in a Web ITS for first-degree equations?
- **RQ2:** How does step-by-step feedback support students after an incorrect answer in the proposed ITS?
- **RQ3:** What implementation and methodological limitations affect mutation-based reasoning reconstruction in this educational context?

2. Literature Review

2.1. Intelligent Tutoring Systems and Error Diagnosis

Intelligent Tutoring Systems (ITS) are computational environments designed to support learning through adaptive interaction, learner monitoring, and personalized feedback. They have been successfully applied in domains such as distance learning, language instruction, and mathematics education, supporting performance prediction, metacognitive development, adaptive problem solving, and the identification of behavioral patterns associated with learning difficulties [Smith et al. 2023, Alam 2023, Crossley et al. 2017, Kautzmann and Jaques 2019, Ruan et al. 2020, Gong and Beck 2015, Xing and Goggins 2015].

A central objective of ITS is to identify student misconceptions and provide appropriate instructional support. Bayesian student modeling, learner profiling, and step-by-step tutoring have demonstrated effectiveness in adapting instruction to learner needs [Gamboa and Fred 2002]. However, these approaches generally detect that an error occurred without explicitly explaining the reasoning process that produced the incorrect answer, particularly in mathematical problem-solving contexts where different misconceptions may lead to similar responses.

To address this limitation, several studies have investigated mechanisms for inferring student reasoning. Brown and Burton [Brown and Burton 1978] introduced procedural bug models to represent recurring mathematical mistakes as incorrect solution strategies rather than isolated errors. Likewise, research on student modeling

and cognitive tutors has shown that representing intermediate reasoning steps can improve diagnosis and enable more personalized instructional support [Azevedo et al. 2013, Chrysafiadi and Virvou 2013].

Despite these advances, most existing approaches rely on predefined bug libraries, cognitive models, or probabilistic representations of learner knowledge. Comparatively fewer studies attempt to reconstruct complete reasoning trajectories capable of reproducing the learner's final answer. This gap motivates the mutation-based approach proposed in this work, in which alternative solution paths generated from atomic error patterns are compared with student responses to produce interpretable hypotheses about the reasoning process underlying observed errors.

2.2. Mutation Analysis and Reasoning Reconstruction

Mutation analysis is a technique originally developed in software testing that introduces controlled modifications into an artifact through mutation operators [Untch et al. 1993]. Originally proposed for fault detection, mutation analysis has subsequently been applied to domains such as deep neural network testing and compiler validation [Wang et al. 2021, Shen et al. 2021].

The underlying principle of mutation analysis can be adapted to educational contexts by representing common misconceptions as controlled modifications of a correct solution process. From this perspective, mathematical errors can be modeled as atomic error patterns associated with specific solution steps. By introducing such patterns into a correct solution, alternative reasoning trajectories can be generated and compared with student responses.

This interpretation extends mutation analysis beyond fault detection and enables the reconstruction of plausible reasoning paths associated with learner answers. Consequently, feedback can be linked not only to an incorrect result but also to the reasoning process that may have generated it [Karpouzis 2023]. Furthermore, mutation-based reconstruction can complement Artificial Intelligence techniques by providing explicit and auditable reasoning trajectories that may support the identification and explanation of candidate misconceptions. Such approaches are particularly relevant in learning environments that require individualized support and interpretable feedback [Costa Santos et al. 2023].

2.3. Related Work

Recent studies have explored the use of Artificial Intelligence to improve personalization, feedback generation, and learner support in Intelligent Tutoring Systems (ITS). A systematic review conducted by Lin et al. [Lin et al. 2023] analyzed the application of AI techniques in ITS and highlighted their potential to support adaptive learning, learner modeling, and personalized educational experiences. Similarly, Létourneau et al. [Létourneau et al. 2025] reviewed AI-driven ITS in K-12 education and observed that most contemporary systems focus on predicting learner performance, adapting instructional content, and recommending learning resources.

In parallel, the increasing use of AI in educational environments has raised concerns regarding transparency and interpretability. Karpouzis [Karpouzis 2023] argues that educational systems should provide explainable diagnostic processes, allowing learners

and educators to understand how decisions and recommendations are generated. Likewise, Türkmen [Türkmen 2025] highlights the growing importance of Explainable Artificial Intelligence (XAI) in education, emphasizing that transparent reasoning mechanisms can improve trust, understanding, and pedagogical value.

Within the Brazilian context, recent studies have investigated the use of generative Artificial Intelligence to support educational activities. Silva et al. [da Silva et al. 2024] discuss how generative AI can assist teaching and learning processes by providing personalized support and adaptive interactions. These studies reinforce the relevance of AI-based educational technologies and demonstrate the growing interest in intelligent systems capable of supporting learners during problem-solving activities.

Although existing ITS research has advanced learner modeling, prediction, personalization, and explainability, few studies explicitly reconstruct reasoning trajectories leading to incorrect answers. The proposed approach addresses this gap by employing mutation analysis to generate alternative solution paths from atomic error patterns, enabling interpretable diagnostic feedback grounded in reconstructed reasoning.

3. A Mutation-Based ITS

This section presents the architecture and operational flow of the proposed Web-based Intelligent Tutoring System (ITS). The system supports learning of first-degree equations by reconstructing plausible reasoning paths associated with student responses. Rather than determining only whether an answer is correct, the ITS attempts to infer how a learner reached that answer by reconstructing alternative solution trajectories from known error patterns.

The architecture follows the traditional ITS structure, comprising pedagogical, domain, student interaction, and interface components. Complementing these modules, a mutation-based reasoning reconstruction engine generates and evaluates alternative solution paths associated with student misconceptions.

The current implementation focuses on first-degree equations. For each interaction, the system generates a problem from a predefined template by randomly modifying the variables involved, ensuring unique instances while preserving the same underlying mathematical structure.

After receiving the student's answer, the domain module generates the correct step-by-step solution, which is forwarded to the mutation engine to reconstruct plausible reasoning paths capable of producing the student's response.

The reconstruction process relies on a knowledge base of atomic error patterns derived from student solutions and consultations with mathematics teachers. These patterns represent recurring misconceptions as elementary error units that can be independently manipulated by the system.

The knowledge base of atomic errors was constructed through an exploratory study conducted prior to the development of the ITS. Written mathematical solutions from 63 students attending Brazilian public schools were analyzed to identify recurring observable errors. The analysis comprised two stages: identification of incorrect solution steps followed by normalization, in which similar errors were grouped into 13 atomic error categories representing the smallest computationally representable mathematical mis-

takes. These categories were subsequently adapted into mutation operators used by the proposed reasoning reconstruction engine.

Examples of atomic errors include sign inversion, arithmetic mistakes, incorrect movement of terms between equation sides, and other elementary transformations frequently observed during equation solving. Although these errors represent observable mathematical actions, they may reflect different underlying causes, including conceptual misunderstandings, procedural misconceptions, inattentiveness, or combinations of these factors.

Once the correct solution has been generated, the mutation engine systematically introduces one atomic error at a time into individual solution steps. Each mutation produces an alternative reasoning trajectory whose final answer is compared with the student's response. This process is repeated for all applicable mutation operators and solution steps, allowing the system to isolate the contribution of each misconception while preserving interpretability.

When a reconstructed trajectory reproduces the student's answer, the associated atomic error is considered a plausible explanation for the observed behavior. Rather than simply classifying the response, the ITS stores the reconstructed trajectory, associates it with one or more candidate misconceptions, and presents these hypotheses as diagnostic information for students or teachers.

The proposed ITS supports two complementary interaction models.

Model A (Reasoning Reconstruction Model).

Model A reconstructs alternative reasoning trajectories from atomic error patterns and identifies candidate misconceptions capable of explaining the student's answer, providing an initial diagnostic feedback.

Model B (Guided Step-by-Step Analysis Model).

Model B presents step-by-step solutions together with distractors derived from the reconstructed trajectories, allowing students to validate, refine, or reject the inferred misconceptions.

The models may operate independently or together, with Model B providing additional evidence to support the diagnostic process. All interactions are logged, including submitted answers, reconstructed trajectories, identified atomic errors, selected distractors, and generated feedback.

For this study, the ITS administered a simulated test composed of ten first-degree equation problems. For each question, the system generated a unique problem instance, reconstructed candidate reasoning trajectories, and provided diagnostic feedback followed by an optional guided interaction.

Figure 1 presents the general operational flow of the proposed mutation-based ITS.

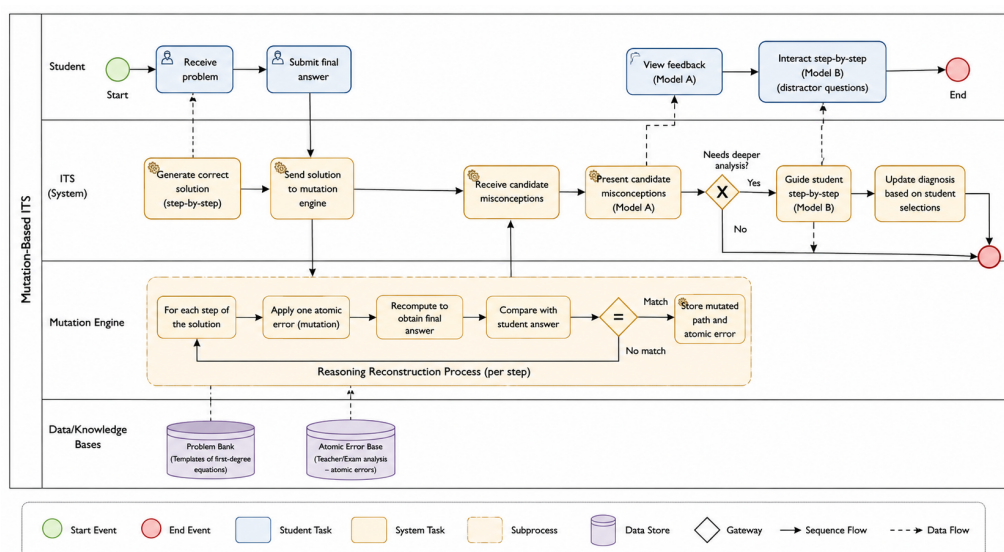


Figure 1. General flow model of the mutation-based ITS.

4. Methodology

This study adopted a quasi-experimental design to evaluate the proposed Intelligent Tutoring System (ITS). According to [Campbell 1963] and [Anderson-Cook 2005], quasi-experiments involve an intervention and the observation of its effects under real-world conditions without random assignment of participants. In this study, participants voluntarily interacted with the Web-based ITS, and their responses were analyzed to investigate the feasibility of mutation-based reasoning reconstruction. Because the study involved an educational intervention without randomization or a control group, it is characterized as quasi-experimental.

The ITS was deployed online, allowing remote interaction. The target audience comprised individuals who had previously studied first-degree equations, including high-school students, university students, workers, retirees, and other participants familiar with the topic. Recruitment was conducted through social media, university classes, public schools, educational projects, and mathematics teachers.

Participation was voluntary and decentralized, allowing users to access the system remotely at their convenience. No personally identifiable or sensitive personal data were collected, and each session received an automatically generated identifier used solely for interaction tracking. According to the Brazilian National Health Council Resolution No. 510/2016, educational studies that do not involve the collection or disclosure of identifiable personal data are exempt from prior review by a Research Ethics Committee. Therefore, this study did not require Research Ethics Committee approval.

The ITS administered ten first-degree equation problems generated from predefined templates. Coefficients and variables were randomized at runtime, producing unique instances while preserving equivalent mathematical structures and difficulty levels. Question order was also randomized for each session.

Participants initially solved each problem by submitting only the final answer. The ITS then generated the correct solution and executed the mutation-based reasoning

reconstruction process. Candidate misconceptions were identified by applying atomic error patterns to intermediate solution steps. When a reconstructed trajectory reproduced the participant's answer, the corresponding misconception was recorded and presented as diagnostic feedback.

Participants could then proceed to a guided step-by-step interaction, in which distractors derived from reconstructed trajectories were presented to validate, refine, or reject the initially inferred misconceptions.

All participants experienced the same experimental condition by interacting with the proposed ITS. The study aimed to evaluate the feasibility of the mutation-based reasoning reconstruction mechanism and analyze participants' interactions with its diagnostic and guided-feedback components, rather than compare instructional approaches. Accordingly, the quasi-experimental design emphasized the empirical evaluation of the proposed educational technology instead of causal inference regarding learning outcomes.

4.1. Data Collection and Preprocessing

All interactions were automatically recorded in activity logs containing submitted answers, identified error patterns, reconstructed trajectories, step-by-step interactions, and generated distractors. After preprocessing, which removed non-essential information and excluded sessions without relevant interactions, the logs were converted into structured datasets containing variables related to student performance, identified misconceptions, mutation matches, guided-interaction outcomes, and aggregated descriptive statistics.

4.2. Data Analysis

The analysis combined descriptive and inferential statistics according to the research questions. RQ1 evaluated the ability of the mutation engine to associate student responses with reconstructed reasoning trajectories. RQ2 compared performance during the initial and guided interaction stages using a Chi-Square test of independence. RQ3 analyzed interaction logs and participant feedback to identify implementation constraints and limitations of the mutation-based reconstruction process. Whenever necessary, aggregated results were validated against the original logs.

Whenever necessary, aggregated results were validated against the original interaction logs to ensure consistency throughout the analysis.

5. Findings

5.1. Results of the Reasoning Reconstruction Process

A total of 12,024 log entries were generated during the experiment, corresponding to 2,004 recorded actions within the ITS. Among 165 recorded accesses, 52 sessions contained no relevant interaction and were excluded from the analysis. Consequently, 113 active participant sessions were considered.

Across these sessions, 594 responses were submitted to first-degree equation problems. Of these, 207 responses (34.8%) were correct on the first attempt, while 387 responses (65.2%) were incorrect.

For each incorrect response, the mutation engine attempted to reconstruct reasoning trajectories using the atomic error patterns available in the knowledge base. Among

the 387 incorrect responses, 70 of the 387 incorrect responses (18.09%) were associated with at least one reconstructed trajectory, while 317 could not be associated with any trajectory generated by the available mutation operators.

Table 1 summarizes the overall results.

Table 1. Results of correct and incorrect answers to the questions

	Total records	Percentage	Std. Dev.	Variance
Correct	207 / 594	34.8%	2.68	7.18
Identified Errors	70 / 594	11.8%	1.32	1.74
Unidentified Errors	317 / 594	53.4%	4.33	18.73

Table 2 presents the distribution of correct responses, identified trajectories, and non-reconstructed responses across the ten question templates.

Depending on the submitted answer, between one and approximately twenty candidate misconceptions were associated with a single response. On average, three candidate trajectories were generated per question.

Table 2. Overview of the results by question

	Correct	Identified	Unidentified	Total
Q1	22	5	34	61
Q2	22	7	31	60
Q3	24	6	30	60
Q4	18	8	33	59
Q5	22	8	29	59
Q6	17	7	35	59
Q7	21	8	30	59
Q8	18	5	36	59
Q9	24	7	28	59
Q10	19	9	31	59
Total	207	70	317	594

5.2. Results of the Guided Step-by-Step Interaction

The second stage of the ITS consisted of a guided step-by-step interaction in which participants analyzed distractors generated from reconstructed trajectories and atomic error patterns.

A total of 133 responses were recorded during this stage. Among them, 101 were correct and 32 were incorrect, corresponding to an accuracy rate of 75.94

To investigate whether response correctness differed between the initial problem-solving stage and the guided interaction, a Chi-Square test of independence was performed using the frequencies presented in Table 3.

The Chi-Square test yielded a statistically significant result ($\chi^2 = 75.15$, $df = 1$, $p < 0.001$), indicating an association between interaction stage and response correctness.

Table 3. Observed frequencies for the Chi-Square test

Stage	Correct	Incorrect
Initial	207	387
Guided	101	32

While 34.8% of the responses submitted during the initial stage were correct (207 out of 594), the guided interaction achieved a correctness rate of 75.94% (101 out of 133).

Table 4 presents the detailed results obtained during the guided interaction.

Table 4. Data from the step-by-step processes

Step / Question	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
Step 1 correct	7	4	3	6	2	4	4	5	10	3	48
Step 1 incorrect	1	7	0	3	1	0	2	0	4	0	18
Step 2 correct	5	2	1	1	3	1	1	3	1	2	20
Step 2 incorrect	2	0	2	1	0	2	0	2	1	0	10
Step 3 correct	3	2	0	1	2	1	1	3	2	1	16
Step 3 incorrect	0	0	0	0	0	1	0	0	0	0	1
Step 4 correct	4	3	0	1	1	0	2	3	1	1	16
Step 4 incorrect	2	0	0	0	0	1	0	0	0	0	3
Step 5 correct	1	0	0	0	0	0	0	0	0	0	1
Step 5 incorrect	0	0	0	0	0	0	0	0	0	0	0
Total	25	18	6	13	9	10	10	16	19	7	133

Most responses were concentrated in the first two reconstruction steps. Step 1 registered 66 interactions (48 correct and 18 incorrect), while Step 2 registered 30 interactions (20 correct and 10 incorrect). The remaining steps accounted for 37 interactions.

6. Discussion

6.1. Discussion of RQ1

RQ1 investigated the extent to which mutation-based reasoning reconstruction can associate student answers with plausible error trajectories.

The mutation engine associated 70 of the 387 incorrect responses (18.09%) with at least one reconstructed trajectory. Although this reconstruction rate appears modest, it should be interpreted in light of the study objective, which was not to maximize error coverage but to evaluate whether mutation operators derived from atomic error patterns could reconstruct plausible reasoning trajectories capable of explaining student answers. Unlike many ITS approaches that emphasize error detection, learner modeling, or performance prediction [Lin et al. 2023, Létourneau et al. 2025, Chrysafiadi and Virvou 2013], the proposed model seeks to generate interpretable hypotheses about the reasoning process itself. Therefore, the observed reconstruction rate primarily reflects the current coverage of the atomic error repository rather than an intrinsic limitation of the mutation-based approach, demonstrating the feasibility of using mutation operators as reasoning

reconstruction mechanisms that complement explainable ITS models [Karpouzis 2023, Lin et al. 2023].

Several factors may explain the large number of unidentified responses. The current implementation introduces only one atomic error at a time, whereas students frequently combine multiple misconceptions within the same solution. In addition, the repository currently contains only thirteen atomic error categories and therefore cannot represent the full diversity of mathematical reasoning strategies. Some responses may also result from arithmetic slips, unconventional solution procedures, or input formatting issues that cannot be reproduced by the available mutation operators. Furthermore, the mutation engine reports a diagnosis only when a complete reconstructed trajectory reproduces the student's final answer. Although this conservative strategy reduces unsupported diagnoses, it naturally limits the overall reconstruction rate. These findings reinforce the need to expand the atomic error repository and investigate multi-mutation reconstruction strategies in future work.

6.2. Discussion of RQ2

RQ2 examined the role of the guided step-by-step interaction following an incorrect answer.

The results indicate a substantial improvement between the initial problem-solving stage and the guided interaction. While only 34.8% of responses were correct during the initial submission, this proportion increased to 75.94% during the guided interaction. The Chi-Square test confirmed a statistically significant association between interaction stage and response correctness ($\chi^2 = 75.15$, $df = 1$, $p < 0.001$), suggesting that participants performed considerably better when interacting with the guided resolution process. These findings are consistent with previous studies showing that guided interactions improve learner engagement and mathematical problem solving in ITS environments [Crossley et al. 2017, Kautzmann and Jaques 2019, Ruan et al. 2020].

Beyond its instructional role, the guided interaction also supports the proposed diagnostic architecture. Because the distractors are derived from reasoning trajectories reconstructed by the mutation engine, students' selections provide additional evidence for validating the inferred misconceptions. The reduction in incorrect responses suggests that participants were generally able to follow the proposed reasoning paths after receiving intermediate guidance, indicating that the guided interaction functions both as a tutoring mechanism and as a validation layer for the diagnostic process.

6.3. Discussion of RQ3

RQ3 addressed the limitations and practical challenges associated with mutation-based reasoning reconstruction.

The main limitation identified was the dependency on the atomic error repository. The reconstruction process can only generate hypotheses from misconceptions represented by the available mutation operators. Consequently, reasoning patterns not modeled in the repository cannot be diagnosed.

Therefore, increasing reconstruction coverage does not necessarily require replacing the mutation-based approach, but rather extending the diversity of modeled misconceptions and allowing combinations of atomic errors during trajectory generation.

This limitation is reflected in the 317 incorrect responses that could not be associated with reconstructed trajectories. These cases may correspond to unmodeled misconceptions, combinations of multiple errors, unexpected reasoning strategies, or input-processing limitations.

Another challenge concerns scalability. As the number of atomic errors increases, the space of possible trajectories grows considerably. Although the current implementation mitigates this issue by introducing a single mutation at a time, larger domains may require mechanisms to prioritize candidate trajectories.

In this context, Artificial Intelligence techniques represent a promising complementary approach. Recent reviews on Explainable Artificial Intelligence (XAI) and AI-driven Intelligent Tutoring Systems have highlighted the potential of AI techniques to support transparent, adaptive, and personalized educational systems [Türkmen 2025, Lin et al. 2023]. Rather than replacing mutation-based reconstruction, machine learning models and large language models may assist in selecting, ranking, and expanding candidate misconceptions. This combination could reduce the search space while preserving the interpretability provided by explicit reasoning reconstruction.

Finally, implementation-related issues, particularly input normalization and response formatting, also affected the reconstruction process. Some responses could not be correctly processed by the prototype, reducing the overall identification rate.

Overall, the findings suggest that mutation-based reasoning reconstruction is a viable and interpretable mechanism for diagnostic feedback in ITS environments. Future work will focus on expanding the atomic error repository through the analysis of additional student solutions, incorporating new categories of mathematical misconceptions, and investigating combinations of multiple atomic errors during trajectory generation. These extensions are expected to increase reconstruction coverage while preserving the interpretability of the proposed mutation-based approach. In addition, hybrid strategies combining mutation analysis with Artificial Intelligence techniques will be investigated to prioritize candidate trajectories and improve the scalability of the reasoning reconstruction process.

7. Threats to Validity

The findings of this study should be interpreted considering some limitations of the experimental design and the implemented prototype.

Regarding internal validity, no control group or pre-test/post-test design was employed. Therefore, although better performance was observed during the guided interaction, the results cannot be interpreted as definitive evidence of learning gains [Létourneau et al. 2025]. Consequently, the quasi-experimental design adopted in this study prioritizes ecological validity and feasibility evaluation over causal inference.

Regarding construct validity, the mutation engine depends on a repository of atomic error patterns. While this repository was derived from student solutions and consultations with mathematics teachers, it does not cover all possible misconceptions. Consequently, some incorrect responses could not be associated with reconstructed reasoning trajectories.

External validity is limited by the voluntary and remote participation model, which

prevented full control over participants' backgrounds and prior knowledge.

Finally, the current implementation focuses exclusively on first-degree equations and includes prototype-specific limitations related to input processing and error categorization. Therefore, the findings should be generalized to other domains only after further empirical validation.

8. Use of Artificial Intelligence Tools

Large Language Model (LLM) tools were used during manuscript preparation exclusively to support grammar correction, language refinement, and text readability improvements. All scientific decisions, methodological choices, analyses, interpretations, and conclusions were performed and validated by the authors. The authors reviewed and approved all manuscript content before submission.

9. Conclusions

This study presented and evaluated a mutation-based reasoning reconstruction mechanism integrated into a mathematics-oriented Intelligent Tutoring System (ITS). The proposed approach uses atomic error patterns to generate alternative reasoning trajectories and associate student responses with plausible misconceptions.

The evaluation involved 113 participants solving first-degree equation problems in a Web-based environment. The results demonstrated that the mutation engine was capable of reconstructing a subset of student errors and generating diagnostic feedback based on plausible reasoning paths. In addition, participants achieved significantly higher performance during the guided step-by-step interaction than during the initial problem-solving stage, suggesting that the proposed interaction model can support reflection and problem understanding.

The main contribution of this work is the reinterpretation of mutation analysis as a reasoning reconstruction mechanism for educational environments. While mutation analysis has traditionally been employed for software quality assurance, fault detection, and software validation [Untch et al. 1993, Wang et al. 2021, Shen et al. 2021], the proposed model adapts mutation operators to generate explainable hypotheses about student reasoning processes and provide interpretable diagnostic feedback in Intelligent Tutoring Systems.

Overall, the findings indicate that mutation-based reasoning reconstruction is a feasible approach for supporting error diagnosis in Intelligent Tutoring Systems and represents a promising direction toward more transparent, explainable, and pedagogically informed educational technologies, complementing current trends in explainable AI for education [Karpouzis 2023, Türkmen 2025].

Future studies will investigate the automatic expansion of the atomic error repository from authentic student solutions, the application of the proposed approach to additional mathematical domains, and comparative evaluations against traditional diagnostic mechanisms adopted in Intelligent Tutoring Systems.

References

Alam, A. (2023). Harnessing the power of ai to create intelligent tutoring systems for enhanced classroom experience and improved learning outcomes. In *Intelligent Commu-*

- nication Technologies and Virtual Mobile Networks*, pages 571–591. Springer Nature Singapore, Singapore.
- Anderson-Cook, C. M. (2005). Experimental and quasi-experimental designs for generalized causal inference.
- Azevedo, R., Aleven, V., et al. (2013). *International handbook of metacognition and learning technologies*, volume 26. Springer.
- Brown, J. S. and Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive science*, 2(2):155–192.
- Campbell, D. T. (1963). and jc stanley. *Experimental and quasi-experimental designs for research*.
- Chrysafiadi, K. and Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, 40(11):4715–4729.
- Costa Santos, L. d., Severo, L. R., and Correia, L. d. A. (2023). Desafios ao engajamento acadêmico no ensino superior: uma análise a partir de posicionamentos sobre satisfação de estudantes. *Revista Internacional de Educação Superior*, 9:e023027.
- Crossley, S., Liu, R., and McNamara, D. (2017). Predicting math performance using natural language processing tools. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*, pages 339–347, New York, NY, USA. ACM.
- da Silva, T. L., Vidotto, K. N. S., Tarouco, L. M. R., and da Silva, P. F. (2024). Potencialidades do uso de inteligência artificial generativa como apoio ao ensino de programação. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1942–1956. SBC.
- de Sousa, H. and Cruz, D. M. (2024). Capacitando educadores com ia generativa: Implicações na educação. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1931–1941. SBC.
- Gamboa, H. and Fred, A. (2002). Designing intelligent tutoring systems: a bayesian approach. In Filipe, J., Sharp, B., and Miranda, P., editors, *Enterprise Information Systems III*, pages 146–152. Springer, New York.
- Gong, Y. and Beck, J. E. (2015). Towards detecting wheel-spinning: Future failure in mastery learning. In *Proceedings of the Second ACM Conference on Learning@Scale (L@S '15)*, pages 67–74.
- Karpouzis, K. (2023). Explainable ai for intelligent tutoring systems. In *International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications*, pages 59–70. Springer.
- Kautzmann, T. R. and Jaques, P. A. (2019). Effects of adaptive training on metacognitive knowledge monitoring ability in computer-based learning. *Computers & Education*, 129:92–105.
- Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., and Léger, P. M. (2025). A systematic review of ai-driven intelligent tutoring systems (its) in k-12 education. *npj Science of Learning*, 10(1):29.

- Lin, C.-C., Huang, A. Y., and Lu, O. H. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart learning environments*, 10(1):41.
- Ruan, S. et al. (2020). Supporting children’s math learning with feedback-augmented narrative technology. In *Proceedings of the Interaction Design and Children Conference*, pages 567–580.
- Shen, Q. et al. (2021). A comprehensive study of deep learning compiler bugs. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE ’21)*, pages 968–980.
- Smith, R. C. et al. (2023). A research agenda for computational empowerment for emerging technology education. *International Journal of Child-Computer Interaction*, 38:100616.
- Türkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. *Journal of Educational Computing Research*, 63(2):277–310.
- Untch, R. H., Offutt, A. J., and Harrold, M. J. (1993). Mutation analysis using mutant schemata. In *Proceedings of the 1993 ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA ’93)*, pages 139–148.
- Wang, Z. et al. (2021). Prioritizing test inputs for deep neural networks via mutation analysis. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 397–409.
- Xing, W. and Goggins, S. (2015). Learning analytics in outer space: a hidden naïve bayes model for automatic student off-task behavior detection. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge (LAK ’15)*, pages 176–183.
- Zerkouk, M., Mihoubi, M., and Chikhaoui, B. (2025). A comprehensive review of ai-based intelligent tutoring systems: applications and challenges. *arXiv preprint arXiv:2507.18882*.