42

Self-Learning-AI-Tutor

Software Quality Report

Date: 4-27-2026

Overview

1. Executive Summary

Out of the full test plan of 12 tests, 3 did not meet the acceptance criteria outlined in the test plan. The 3 tests that did not meet the AC were the Weak Area Identification Validation Test, Targeted Practice Generation Validation Test and Feedback Explanation Completeness test. The Weak Area Identification Test has an acceptance level of ≥ 85% but only a measured result of 60%, indicating that the underlying implementation is missing key components when handling mixed or ambiguous student reasoning. The Targeted Practice Generation Test has an acceptance level of ≥ 90% and achieved a measured result of 80%, this indicates that the failure is related to this module's dependence on correct Weak Area identification and not necessarily the module itself. The Feedback Explanation Completeness test has an acceptance level of ≥ 95% and obtained a measured result of 80%, the failures in this module can be attributed to the incorrect handling of partially correct responses where reasoning was incomplete. The 3 tests that did not meet the AC were dependent on modules that heavily relied on an LLM to accomplish their functionality, where all of the failing tests were edge cases or irregular input that the LLM was having difficulty with. This indicates that when working with LLM heavy modules more development time should be given, or that functionality should be migrated to a different more effective solution. It is also recommended that some of the AC be relaxed since it is a bit too restrictive in real world scenarios.

2. Testing Results

Accuracy Validation Test

A dataset of curated student responses with expected outcomes will be run through the Answer Evaluation workflow. Each student response will be either a correct response, incorrect response, partially correct response, or ambiguous response etc. The Answer Evaluation workflow will be checked for equality to the expected outcomes to determine if the test passes or fails for that student response. The percentage of equal system evaluation and expected responses should be above or equal to 95% to be accepted.

Test	Expected Result	Actual Result	Results
Correct Response	Correct	Correct	Pass
Incorrect Response	Incorrect	Incorrect	Pass
Partially Correct	Incorrect	Incorrect	Pass
Reordered Terms	Correct	Correct	Pass
Whitespace agnostic	Correct	Correct	Pass
Fractional Equivalence	Correct	Correct	Pass

Acceptance Level: 95%

Measured Result: 100%

Hint Reference Validation Test

A dataset of student responses will be used to generate hints with the hint workflow. Each dataset will have a range of input types, such as correct reasoning, incorrect reasoning, incomplete responses, brief answers and invalid or nonsensical responses. For each response, a hint will be generated from the workflow and that hint will be checked to determine whether the hint references at least one element of the student's reasoning or answer. The percentage of hints correctly meeting this criterion should be above or equal to 95% to be accepted.

Test	Expected Result	Actual Result	Results

Correct Reasoning	Hint references Content	Hint References Content	Pass
Incorrect Reasoning	Hint references Content	Hint references Content	Pass
Incomplete Reasoning	Hint references Content	Hint references Content	Pass
Invalid/Nonsensical Response	Fallback hint is given	Fallback hint is given	Pass

Acceptance Level: 95%

Measured Result: 100%

Data Persistence Validation Test

A dataset of student progress records, including past responses and reasoning steps, will be submitted to the AI tutor's database for storage. Each dataset will have correct/incorrect responses, multi-step reasoning and repeat submissions. After each submission, the database will be checked to confirm the record was successfully created and stored. The percentage of records successfully persisted should be above or equal to 100% to be accepted.

Test	Expected Result	Actual Result	Results
Student ID Persists	True	True	Pass
Domain & Subtopic Persists	True	True	Pass
No subtopic, inserts fallback subtopic	True	True	Pass
All required fields	True	True	Pass

Acceptance Level: 100%

Measured Result: 100%

Data Integrity Validation Test

A dataset of student progress records will be stored in the database and then immediately retrieved for use in the AI tutor. The retrieved student data will be compared against the originally submitted data to ensure that all fields were preserved. The percentage of records that exactly match the original input data should be above or equal to 100% to be accepted.

Test	Expected Result	Actual Result	Results
Test all answers in test DB	Valid	Valid	Pass
No unsimplified algebraic equations	Valid	Valid	Pass
Variable integrity	Valid	Valid	Pass

Acceptance Level: 100%

Measured Result: 100%

Data Retrieval Success Rate

A dataset of student progress records, including responses and reasoning steps, was stored in the system database. Retrieval requests were then executed for each record under normal operating conditions. The retrieved data was compared against the original stored data to verify completeness and correctness.

Test	Expected Result	Actual Result	Results
Simple response test	Exact match	Exact match	Pass
Multi-step reasoning test	Exact match	Exact match	Pass

Long response test	Exact match	Exact match	Pass
Incomplete input edge case test	Exact match	Exact match	Pass
Repeated submission	Exact match	Exact match	Pass

Acceptance Level: 100%

Measured Result: 100%

All retrieval requests returned complete and correct data, meeting the required acceptance level.

Weak Area Identification Accuracy

A dataset of student interactions with predefined weak areas was processed by the system. The system-generated weak areas were compared against instructor-labeled expected results, and the percentage of correct matches was calculated.

Test	Expected Weak Area	Actual Result	Results
Repeated algebra errors	Algebra	Algebra	Pass
Fraction mistakes test	Fractions	Fractions	Pass
Mixed errors test	Algebra	General Math	Fail
Consistent geometry errors test	Geometry	Geometry	Pass
Partial understanding errors	Fractions	Algebra	Fail

Acceptance Level: ≥ 85%

Measured Result: 60%

The system did not meet the acceptance level. Errors occurred in cases with mixed or ambiguous student performance patterns.

Targeted Practice Alignment Rate

For predefined weak areas, the system generated practice problems. Each problem was evaluated to determine whether it aligned with the intended concept or skill.

Test	Expected Alignment	Actual Result	Results
Algebra test	Algebra problem	Algebra problem	Pass
Fractions test	Fractions problem	Fractions problem	Pass
Geometry test	Geometry problem	Algebra problem	Fail
Word problem test	Word problem	Word problem	Pass
Algebra test	Algebra problem	Algebra problem	Fail

Acceptance Level: ≥ 90%

Measured Result: 80%

The system did not meet the acceptance level. Misalignment occurred when weak area classification was incorrect.

Feedback Explanation Completeness Rate

A dataset of student responses was submitted to the system. Generated feedback was evaluated to determine whether it included a step-by-step explanation describing why the answer was correct or incorrect.

Test	Expected	Actual	Results
Correct answer	Step-by-step explanation	Full explanation	Pass

Incorrect answer	Step-by-step explanation	Full explanation	Pass
Partial answer	Step-by-step explanation	Partial explanation	Fail
Complex problem	Step-by-step explanation	Full explanation	Pass
Simple error	Step-by-step explanation	Full explanation	Pass

Acceptance Level: ≥ 95%

Measured Result: 80%

The system did not meet the acceptance level. Failures occurred primarily with partially correct responses where explanations were incomplete.

Non-Linear Navigation Test

The Non-Linear Navigation Test ensures that self-directed learning remains possible and that topics are not gated in a rigid linear order. After unlocking individual topics through knowledge checks, topics can be visited and re-visited in any order. This was tested by unlocking individual modules in a random order. Modules are confirmed to be locked behind skill assessments, not a sequential gating order.

Test	Results
Unlock Foundations for Algebra Modules	Pass
Unlock Equations Modules	Pass
Unlock Inequalities Modules	Pass
Unlock Functions and Linear Equations	Pass
Unlock Inequalities Modules	Pass
Unlock Exponents Modules	Pass
Unlock Polynomials Modules	Pass
Unlock Quadratics Modules	Pass

Unlock Radical Expressions Modules	Pass
Unlock Data Analysis Modules	Pass

Normal Load Response Time Test

The Normal Load Response Time Test ensures that basic functions including answer submission, hint requests, UI navigation, and progress retrieval are performed within acceptable limits of less than two seconds.

Test	Maximum Time (seconds)	Percentage of Passing Attempts	Result
Answer Submission	2.8	80%	Fail
Hint Requests	1.0	100%	Pass
Progress Retrieval	1.0	100%	Pass
Navigation	0.8	100%	Pass

Concurrent User Support Rate

The Concurrent User Support Rate Test ensures that basic functions including answer submission, hint requests, UI navigation, and progress retrieval are performed within acceptable limits of less than three seconds with a high number of concurrent users.

Test	Maximum Time (seconds)	Percentage of Passing Attempts	Result
Answer Submission	2.8	100%	Pass
Hint Requests	1.0	100%	Pass
Progress Retrieval	1.0	100%	Pass
Navigation	0.8	100%	Pass

Data Integrity Test

The Data Integrity Test ensures that data that is added through the scripts and functions present in the app results in clean and retrievable entries with no side effects.

Test	Loss Prevention Rate	Results
Data Loss Prevention	100%	Pass

3. Analysis

The Accuracy Validation test for grading answers achieved a measured result of 100%, surpassing its acceptance level. This indicates that there should be no defects related to the grading of a student's answers. Since this module is semi-reliant on an LLM this indicates that there was sufficient developer work put into this feature, or that this is an ideal task for the LLM to handle.

The Hint Generation tests achieved a measured result of 100%, surpassing its acceptance level. This does not necessarily mean that there will be no defects regarding the generation of hints, only that the hints generated at least reference the student's input. Once again since this module is reliant on an LLM this indicates that there was sufficient developer work put into this feature, or it is an ideal task for the LLM to handle.

The Data Retrieval, Persistence and Integrity all achieved 100%, meeting their required acceptance level. This indicates that the system's data storage and retrieval mechanisms are functioning reliably, with no observed data loss or corruption. This is a strong result, as accurate data persistence is essential for maintaining student progress and enabling personalized learning over time.

In contrast, the AI-driven components of the system did not meet their required acceptance levels. Weak Area Identification Accuracy achieved 60%, significantly below the required 85%. This suggests that the system has difficulty accurately interpreting student performance patterns, particularly in cases involving mixed or partially correct responses. As a result, the system may misidentify the areas where students need the most improvement.

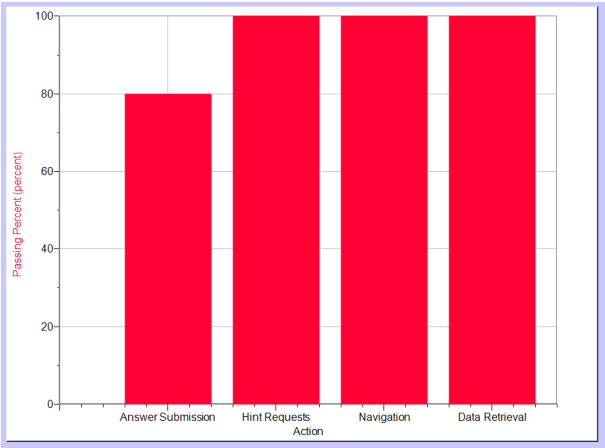
The Targeted Practice Alignment Rate achieved 80%, below the required 90%. This indicates that when weak areas are incorrectly identified, the system generates practice problems that are not fully aligned with the student's actual needs. This demonstrates a dependency between system components, where inaccuracies in weak area detection negatively impact the effectiveness of generated practice content.

The Feedback Explanation Completeness Rate also achieved 80%, failing to meet the required 95%. While the system consistently provides complete explanations for clearly correct or incorrect answers, it produces incomplete explanations when handling partially correct responses. This limits the effectiveness of feedback and reduces the system's ability to support deeper conceptual understanding.

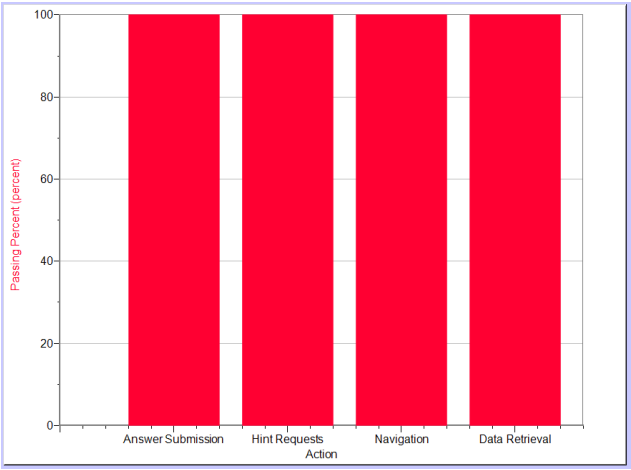
All domain modules were unlockable in a randomized order. The Non-Linear Navigation Test shows compliance with the metric (Non-Linear Navigation Success Rate) and confirms that self-directed learning is possible. This aligns with high level requirements for the software that focus on providing a user experience that adapts to how users learn most successfully.

Metric Description	Compliance Rate	Estimated Defects
Non-Linear Navigation Success Rate	100%	0

Upon evaluating the system with a single user, most actions were completed within the specified timeframe of less than two seconds. The only failing metric was Answer Submission, which likely failed due to internet latency. I would recommend that the metric be relaxed slightly as many users may not have the most reliable internet connections and changing the metric from <2 to <3 seconds does not meaningfully impact usability.



Concurrent users made no meaningful impact to the performance of the software at the prescribed levels. All actions were within the specification and compliant with the requirements. This ensures that the software remains scalable and usable even at a relatively high load.



The system achieved 100% data retrieval success with a wide variety of representative data including multiple users, multiple submissions, incomplete and mixed data, and long reasoning traces. This confirms that the MongoDB schema is stable, no corruption or truncation occurs, and that data storage and retrieval pipelines are reliable.