Enterprise Knowledge-Grounded Human–AI Collaborative Feedback Generation Using Retrieval-Augmented Models

Hanum Khairana Fatmah[1,*], M Itmamul Wafa[2]

[1,2]*Magister of Computer Science, Universitas Gadjah Mada, Indonesia*

(Received: January 1, 2026; Revised: February 28, 2026; Accepted: April 13, 2026; Available online: July 2, 2026)

Abstract

Adaptive feedback generated by large language models often suffers from limited auditability and inconsistent pedagogical intent, which constrains trust in real learning deployments. This study proposes an explainable adaptive feedback framework that combines retrieval-augmented generation with pedagogical rationale tracing, linking learner-state signals, retrieved evidence, and instructional move sequencing into a traceable rationale artifact. Across four conditions, the full model improves faithfulness to evidence from 0.66 to 0.88 and reduces scope violations from 7.4% to 1.9%. Learning proxies improve concurrently, with next-item accuracy increasing from 71.2% to 78.9% and revision uptake rising from 34.6% to 47.8%. Human rubric scores confirm higher instructional usefulness, where actionability increases from 3.6 to 4.3 on a 5-point scale while tone remains stable from 4.1 to 4.2. Ablation results show retrieval as the dominant driver of grounding, with faithfulness dropping to 0.62 and scope violations rising to 8.1% when retrieval is removed, whereas removing rationale tracing mainly degrades actionability from 4.3 to 3.7 and revision rate from 47.8% to 39.5%. Stratified analysis indicates the strongest benefits for low mastery and high frustration learners, where next-item accuracy improves from 61.0% to 70.0% and revision rate increases from 45.0% to 58.0%, alongside persistence gains from 68.3% to 79.5%. Mitigation controls further reduce evidence mismatch from 9.6% to 4.1% and rationale incoherence from 6.8% to 2.9%. The findings indicate that grounded generation and explicit pedagogical rationale tracing jointly improve effectiveness, accountability, and deployment readiness of adaptive feedback systems.

Keywords: Adaptive Learning, Retrieval-Augmented Generation, Explainable AI In Education, Pedagogical Rationale Tracing, Learner Modeling, Grounded Feedback, Educational Data Mining

1. Introduction

Generative AI has rapidly entered digital learning ecosystems, where large language models are used for tutoring, feedback generation, and assessment support. Empirical evidence suggests benefits for engagement and learning efficiency, yet it also highlights reliability risks when feedback is inconsistent, unverifiable, or poorly aligned with instructional intent [1]. Systematic evidence on pedagogical agent communication further indicates that educational effectiveness depends on adaptive, relational, and logically structured feedback, which current generative systems still struggle to sustain at scale [2].

Adaptive learning systems rely on feedback that is timely, individualized, and instructionally calibrated, because feedback functions as the primary control signal that shapes learner trajectories. Systematic reviews of feedback in intelligent tutoring systems emphasize that effectiveness depends on precision, modality, and alignment to learner state rather than generic correctness statements [3]. Recent adaptive tutoring implementations in STEM show measurable gains from real-time personalization, but also expose a recurring limitation: learners and instructors cannot easily audit why a specific intervention was selected, reducing trust and adoption [4].

A central technical barrier is the epistemic fragility of feedback generated purely from parametric model memory. Retrieval-augmented generation addresses this limitation by conditioning generation on external evidence, enabling provenance and reducing hallucination in knowledge-intensive tasks [5]. However, surveys of retrieval-augmented large language models show that retrieval alone does not guarantee pedagogical usefulness, because evidence relevance

***Corresponding author: Hanum Khairana Fatmah (hanumkhairanafatmah@ugm.ac.id)**
DOI: https://doi.org/10.47738/ijaim.v6i2.122
This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

is necessary but not sufficient for choosing the right instructional move, tone, and granularity for a given learner state [6].

Educational settings impose constraints that intensify these issues, including curricular scope boundaries, misconception-specific guidance, and the need to justify feedback decisions for accountability. A systematic survey of RAG in education reports that hallucination mitigation and knowledge updates improve factuality, but feedback still fails when retrieval does not reflect instructional intent or when explanations are not explicitly tied to learner needs [7]. Work on grounding behavior further suggests that models can produce fluent responses that appear cooperative while still lacking grounded acts that support mutual understanding and learning progress [8].

Evaluation and monitoring remain challenging because feedback quality is multidimensional and cannot be captured by a single accuracy metric. RAG pipelines require reference-free evaluation across retrieval quality, faithfulness, and generation utility, motivating structured assessment frameworks for continuous iteration [9]. Broader NLG surveys similarly emphasize that modern generation systems must be evaluated for controllability, factuality, and user-centered utility, particularly when text functions as guidance rather than information [10]. These issues are amplified in adaptive feedback, where errors can redirect learning paths.

Explainability provides a complementary pathway, not as an optional transparency layer, but as a mechanism for stable decision support in learning analytics and adaptive interventions. Explainable learning analytics demonstrates that explanation methods can reveal stability, drift, and cohort-level consistency, which are essential for operational governance in educational environments [11]. Interpretable machine learning reviews further show that explanation quality must be evaluated against human needs, because explanations that are technically correct can still be unusable for teachers and learners if they do not support actionable decisions [12].

This paper addresses the gap between grounded generation and pedagogically meaningful explainability by proposing Explainable Adaptive Feedback Generation with retrieval-augmented models and pedagogical rationale tracing. The novelty lies in treating pedagogical rationale as a traceable decision object that links learner state signals, retrieved evidence, and feedback move selection, enabling auditability and systematic error diagnosis. This approach is aligned with rationalization research in explainable NLP, which frames natural language rationales as accessible explanations that can be evaluated for faithfulness and coherence [13], and it is motivated by evidence that educational feedback systems require reliability and consistency checks to be trusted as evaluators and feedback providers [14].

2. Literature Review

Contemporary adaptive learning increasingly relies on generative models to deliver formative feedback, yet the pedagogical value of such feedback depends on transparency, learner controllability, and alignment with instructional intent. Explainable AI in education emphasizes that explanations should support actionability for learners and educators, not only post hoc interpretability for developers, and this requirement becomes sharper when feedback is produced in natural language rather than as scalar predictions. Open learner model research further shows that making learner state visible can shape self-regulation behaviors and help calibrate engagement [15], [16].

A second line of work concerns learner-state inference as the substrate for adaptive feedback. Knowledge tracing has progressed from recurrent models to attention-based architectures that better capture long-range dependencies in sparse interaction sequences while maintaining usable predictive calibration. Recent deep knowledge tracing variants address data sparsity through structured attention and generative decoding, indicating that adaptation quality is constrained by both the fidelity of latent mastery estimates and the stability of predictions under limited evidence. These findings motivate explicit uncertainty and reliability signals in downstream feedback generation [17].

In parallel, retrieval-augmented generation has been positioned as a practical mechanism for grounding educational dialogue systems in curated instructional content and institutional policies. Surveys of RAG-based educational chatbots report strong adoption for course support, tutoring-style question answering, and formative assessment, with hallucination reduction as the dominant motivation. However, benchmark studies show that RAG performance is bottlenecked by abilities such as negative rejection and robustness to distractor documents, implying that pedagogical feedback quality is sensitive to retrieval noise and corpus mismatch [18], [19].

Explainability research for large language models adds a complementary perspective by distinguishing transparency of internal computations from explanation quality as experienced by end users. Recent surveys argue that LLM explainability should be treated as a socio-technical problem involving faithful attribution, communicative adequacy, and evaluation protocols that connect explanations to user decisions. For adaptive feedback, this literature implies that "explainable" must cover not only why a response was produced, but also why it is instructionally appropriate given learner state, task constraints, and evidence provenance [20].

Faithfulness remains a central technical concern because plausible explanations can still misrepresent the true determinants of model outputs. Natural language explanation work proposes targeted tests, including counterfactual editing and input reconstruction, to detect unfaithful rationales that do not track decision evidence. Such testing is directly relevant to pedagogical rationale tracing, where a feedback message should be supported by retrievable evidence and consistent with learner modeling signals; otherwise, explanations risk becoming persuasive narratives without instructional validity [21], [22].

Finally, education-specific evidence highlights that explainability in learning analytics and EDM is still unevenly evaluated, with many studies reporting model performance while under-specifying explanation quality metrics and stakeholder outcomes. Umbrella reviews identify a gap in standardized evaluation of explanation usefulness, despite increasing deployment complexity. In response, self-rationalizing approaches that ground predictions and rationales in external knowledge have emerged, providing a conceptual bridge between grounded retrieval and explanation generation that is compatible with pedagogical rationale tracing [23], [24].

3. Methodology

3.1. Learning Context, Data Pipeline, and State Representation

The study operationalizes adaptive feedback in a modular learning environment where each interaction produces fine-grained traces of comprehension, effort, and pacing. Logs include item attempts, response latency, hint usage, revision events, and post-feedback actions, aggregated at session and topic levels. A learner state vector is computed after each item, enabling temporally consistent adaptation. The pipeline enforces strict time-ordering to prevent leakage between prior and subsequent feedback events [3], [17].

$$s_t = \phi(s_{t-1}, x_t) \tag{1}$$

The state update function $\phi(\cdot)$ fuses the previous state st-1 with new evidence xt from the current interaction, producing s_t. In practice, $\phi(\cdot)$ is instantiated as a gated recurrent update with calibrated feature scaling, so that high-variance signals such as latency do not dominate mastery indicators. This formulation supports incremental adaptation while preserving stability across short bursts of noisy behavior [4].

Figure 1 visualizes the full preprocessing path that converts raw learner interaction events into time-ordered features and an incremental learner state used by downstream adaptation. The top flow emphasizes leakage prevention by enforcing strict chronological ordering before constructing windowed features and state updates. The lower flow clarifies how trigger conditions define feedback opportunities and how outcome labels are aligned to those triggers. This structure makes the adaptation decision traceable to specific observed behaviors and measurable post-feedback outcomes.

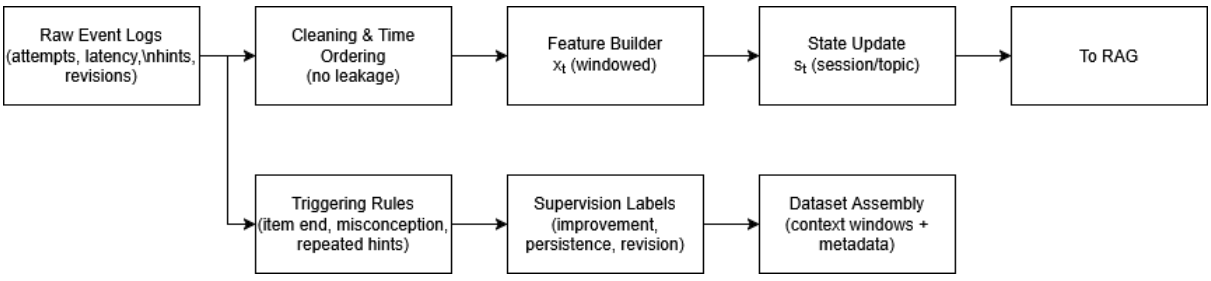

Figure 1. End-to-end Pipeline from Logs to Learner State and Supervision

Table 1 specifies the minimal feature set that operationalizes learner behavior for adaptation, separating item-level signals from session and topic aggregates to preserve temporal granularity. The window column enforces consistency in how evidence is accumulated, which is critical for comparing models under identical observational constraints. The expected effect column encodes interpretive directionality used both for prompt control and later explanation generation, so that feedback decisions can be audited against concrete behavioral signals rather than implicit model heuristics.

Table 1. Feature Schema for x_t and State Indicators

Feature	Level	Definition	Window	Expected Effect	Type
N_attempts	Item	Number of attempts on current item	Current	Higher implies difficulty or misconception risk	Count
T_latency	Item	Response time from item open to submit	Current	Higher implies cognitive load or uncertainty	Seconds
N_hints	Item	Total hints requested before final answer	Current	Higher implies support need	Count
Revise_flag	Item	Whether learner revised an answer after feedback	Post	1 indicates actionability and engagement	Binary
Acc_3	Session	Mean accuracy over last 3 items	Last 3	Lower implies remediation priority	Ratio
Persist_5	Session	Completion rate over last 5 triggered items	Last 5	Lower implies motivational scaffolding	Ratio
Topic_mastery	Topic	Calibrated mastery estimates for current topic	Running	Lower implies conceptual clarification	Score
Frustration_proxy	Session	Composite of retries, latency spikes, and hint bursts	Last 10 min	Higher implies tone-softening and stepwise hints	Score

The resulting dataset is aligned to feedback opportunities by defining trigger points at item completion, misconception detection, or repeated hint requests. Each trigger is paired with a context window containing recent interactions and relevant curricular metadata. Labels include short-term improvement, persistence, and corrective revision, enabling multi-objective evaluation. The representation is designed to be model-agnostic, so retrieval and generation modules can be compared under identical state inputs [16].

3.2. Retrieval-Augmented Feedback Generation

The feedback generator follows a retrieval-augmented generation architecture that grounds responses in vetted pedagogical resources and prior validated feedback exemplars. A hybrid retriever combines dense semantic similarity with sparse keyword matching to balance conceptual alignment and terminology precision. Retrieved passages are constrained by curriculum scope and difficulty band derived from the learner state. The generator receives a structured prompt containing learner state, task context, and retrieved evidence [5], [6].

$$z = \sum_{i=1}^{k} \alpha_i d_i, \qquad \alpha_i = \frac{exp\ exp\ (sim(q, d_i)/\tau)}{\sum_{j=1}^{k} exp\ (sim(q, d_j)/\tau)} \tag{2}$$

The evidence aggregation vector z is computed from retrieved documents d_i using temperature-scaled attention weights α_i. The similarity function sim $(\cdot)$ is implemented as cosine similarity between normalized embeddings, while τ controls retrieval sharpness. Lower τ increases concentration on top-ranked evidence, which improves factual grounding but can reduce diversity of pedagogical strategies [18].

Figure 2 formalizes the feedback generator as a grounded architecture that separates retrieval from generation to improve factuality and pedagogical consistency. Parallel dense and sparse retrievers capture conceptual similarity and terminology alignment, then a filter enforces curriculum scope and difficulty bands derived from learner state. The lower pathway highlights that decoding is governed by an explicit policy controlling tone and length, ensuring that adaptivity is not only semantic but also communicative and affect-sensitive.

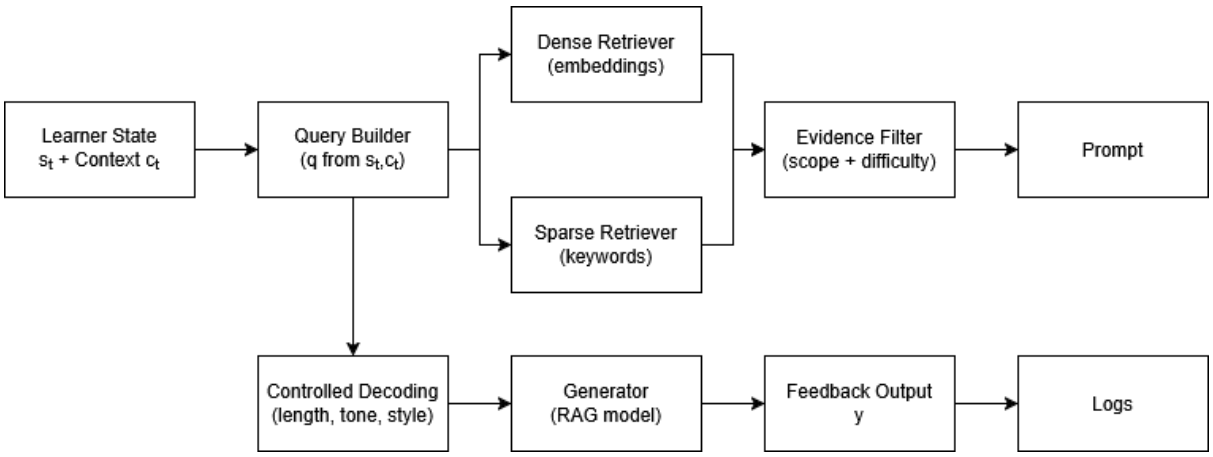

Figure 2. Retrieval-Augmented Feedback Generation with Hybrid Retrieval and Control Policy

Table 2 documents the operational settings that make the RAG pipeline auditable, connecting each configuration choice to a pedagogical purpose and an observable failure signal. This structure supports reproducibility because retrieval and decoding are frequent sources of hidden variance across runs and deployments. By specifying output artifacts and failure indicators, the table also defines what is logged for later analysis, enabling targeted debugging when feedback is off-scope, mismatched in difficulty, or stylistically inappropriate.

Table 2. Retrieval and Decoding Configuration

Component	Setting	Purpose	Constraint Source	Output Artifact	Failure Signal
Dense Retrieval	Top-k=8	Conceptual grounding	Embedding similarity	Ranked passages	Low similarity concentration
Sparse Retrieval	Top-k=8	Terminology match	Keyword overlap	Ranked passages	Missing key terms
Hybrid Merge	Weighted union	Balance recall and precision	Fixed weights	Merged candidate set	Redundant evidence
Scope Filter	Curriculum tags	Prevent off-syllabus content	Course metadata	Scoped evidence	Scope violation count
Difficulty Filter	Band by mastery	Match cognitive demand	Topic_mastery	Leveled evidence	Over-challenging rate
Decoding Policy	Max tokens + tone	Control length and affect	Frustration_proxy	Feedback text	Excess verbosity or harsh tone

To ensure adaptive precision, generation is constrained by a rubric that encodes feedback intent, such as corrective explanation, metacognitive prompt, or motivational reframing. Decoding uses length and style controls tied to state indicators like cognitive load and frustration risk. The prompt template explicitly binds claims to retrieved spans, enabling later attribution. This design reduces hallucination while keeping the feedback contextually aligned with the learner's immediate needs [7].

Pseudo-code 3.1: Retrieval-Augmented Adaptive Feedback

Input: learner state s_t, task context c_t, query builder g (·), index I, generator G
1: q ← g (s_t, c_t)
2: D ← HybridRetrieve(I, q, k)
3: D ← FilterByScopeAndDifficulty(D, s_t, c_t)
4: P ← BuildPrompt(s_t, c_t, D)
5: y ← ControlledGenerate(G, P, constraints = Policy(s_t))
6: return y, D

3.3. Pedagogical Rationale Tracing

Pedagogical rationale tracing represents the underlying instructional justification of a generated feedback message as an explicit, queryable structure. Rationales are modeled as a directed acyclic graph whose nodes denote pedagogical moves, such as error diagnosis, concept linkage, worked-example hinting, or self-explanation prompts. Edges encode dependency relations, ensuring that motivational framing does not contradict correctness explanations. Each generated feedback is mapped to a rationale graph through a constrained extraction step [15], [24].

$$R = (V, E), \qquad E = \{(v_i, v_j) | v_i \rightarrow v_j\} \tag{3}$$

The rationale graph R is defined by a node set V and edge set E indicating precedence constraints between pedagogical moves. Inference enforces a minimality prior, preferring the smallest graph that explains the message while covering required instructional components. This structural constraint supports consistent pedagogical sequencing, for example requiring diagnosis before remediation and requiring remediation before transfer prompts [13].

Figure 3 illustrates pedagogical rationale tracing as an explicit dependency graph that sequences instructional moves from trigger evidence to diagnosis, clarification, and action-oriented guidance. The graph supports explainability because each node corresponds to a pedagogically interpretable function, while edges impose coherence constraints such as requiring diagnosis before remediation and placing transfer checks after actionable steps. This representation enables post-hoc auditing of whether generated feedback follows a defensible instructional logic rather than an opaque generative pattern.

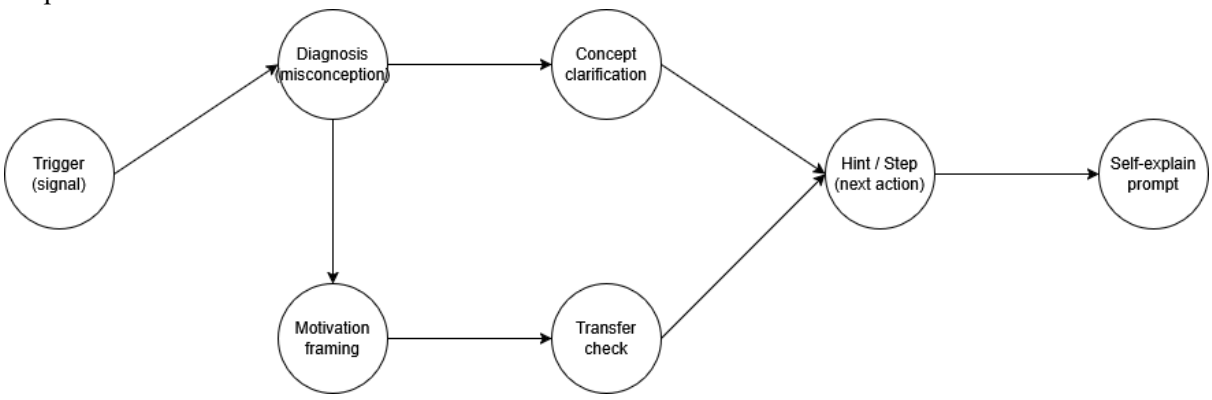

Figure 3. Pedagogical Rationale Tracing as A Directed Graph of Instructional Moves

Table 3 defines a compact taxonomy that bridges instructional theory and implementable detection logic by pairing each pedagogical move with a primary trigger and a concrete evidence link. The "expected output" column anchors each move to an observable feedback segment, which improves annotation consistency and supports automated verification. The "quality risk" column anticipates systematic failure modes, enabling targeted evaluation of rationale coherence and preventing common degradations such as generic encouragement or over-explanatory hints.

Table 3. Pedagogical Move Taxonomy and Tracing Signals

Move	Operational Meaning	Primary Trigger	Evidence Link	Expected Output	Quality Risk
Diagnosis	Identify likely error source or misconception	Repeated incorrect attempts	Error pattern + retrieved exemplar	Short causal explanation	Overconfident mislabeling
Concept clarification	Restate target concept at matched difficulty	Low topic mastery	Curriculum passage	Focused explanation	Excess abstraction
Hint / next step	Provide minimal actionable guidance	High hint usage or long latency	Worked-step snippet	Single next action	Giving full solution
Self-explain prompt	Ask learner to articulate reasoning	Surface-level correct but unstable	Concept checkpoint	Targeted question	Vague prompt

Motivation framing	Normalize struggle and maintain engagement	High frustration proxy	State features	Supportive tone	Overly generic reassurance
Transfer check	Verify generalization to new condition	Post-correction readiness	Related item metadata	Quick follow-up test	Premature escalation

Rationale tracing also records evidence provenance by linking each rationale node to retrieved spans and learner-state signals that triggered the move. For instance, repeated incorrect attempts activate a diagnosis node tied to misconception patterns, while long latency activates a cognitive-load mitigation node. This linkage enables later audits of whether the feedback is justified by observable behavior and supported content. The rationale representation serves as the central interface for explainability and error analysis [11].

3.4. Explainability Outputs and Evaluation Protocol

Explainability is operationalized as a multi-view artifact consisting of attribution, rationale graphs, and counterfactual sensitivity summaries. Attribution reports identify which retrieved spans and state features most influenced each feedback segment. Rationale graphs provide a pedagogical narrative that aligns with instructional theory, while counterfactuals show how feedback would change under alternative learner states. Evaluation measures combine learning impact, explanation faithfulness, and alignment with pedagogical standards [12], [20].

$$Faith(y) = 1 - \frac{1}{|y|}\sum_{t=1}^{|y|} 1[\neg Supported(y_t, D)] \quad (4)$$

Faithfulness is computed as the proportion of generated units y_t that are supported by retrieved evidence D, where Supported $(\cdot)$ is a verifier that checks semantic entailment or strict citation linking. Higher scores indicate reduced unsupported claims. This metric is paired with adaptivity metrics that measure alignment between feedback intent and learner-state needs, ensuring that grounded outputs remain pedagogically responsive [9].

Figure 4 summarizes the evaluation protocol as a triangulated pipeline that links generated feedback and retrieved evidence to automated checks, grounding verification, and rubric-based human review. The lower branch explicitly incorporates behavioral proxies such as revision and persistence, preventing explainability metrics from being treated as purely cosmetic. The final triangulation stage emphasizes agreement analysis across metrics and error inspection, which is essential when improvements in engagement can coexist with reduced faithfulness or inflated verbosity.

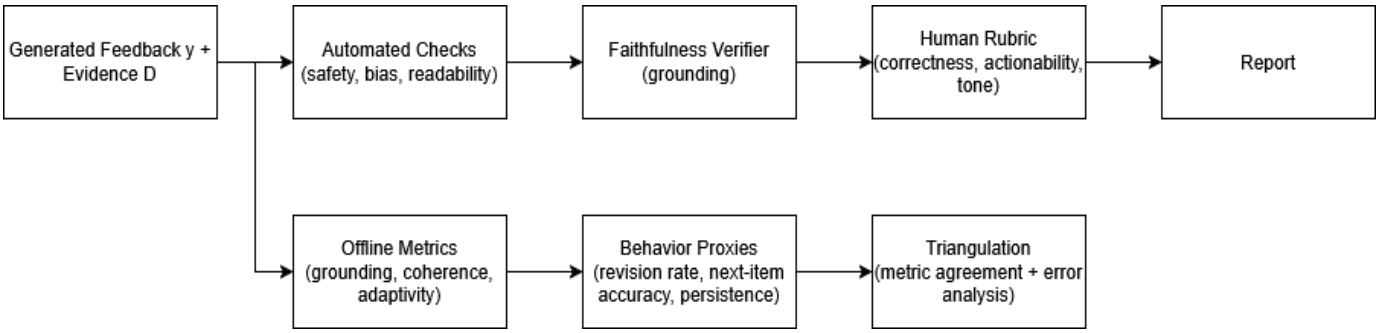

Figure 4. Evaluation Protocol Combining Automated Verification, Human Rubrics, and Behavior Proxies

Table 4 consolidates the core evaluation signals into a balanced suite that jointly measures grounding, pedagogical quality, and downstream behavioral effects. The table also encodes the provenance of each metric, separating automated verification from human judgment and interaction-log outcomes, which prevents conflating interpretability with instructional effectiveness. By listing common failure cases, the table functions as a diagnostic map that guides targeted audits, such as distinguishing genuine revision uptake from superficial edits.

Table 4. Metric Suite and Operationalization

Metric	Target Property	Unit	Measured From	Interpretation	Common Failure
Faithfulness	Grounding to evidence	0-1	Verifier over y and D	Higher means fewer unsupported claims	Paraphrase not linked
Adaptivity Match	Alignment to learner needs	0-1	Intent vs state rubric	Higher means better state-conditioned intent	Generic feedback
Actionability	Clarity of next step	1-5	Human rubric	Higher means concrete and feasible guidance	Overly abstract advice
Tone Appropriateness	Affect suitability	1-5	Human rubric	Higher means supportive and calibrated tone	Harsh or patronizing phrasing
Revision Rate	Post-feedback engagement	%	Logs after feedback	Higher indicates uptake of guidance	Gaming via trivial edits
Next-Item Accuracy	Short-term learning proxy	%	Next assessment item	Higher suggests immediate understanding gain	Item difficulty mismatch

Human evaluation uses a rubric capturing correctness, clarity, tone appropriateness, actionability, and pedagogical coherence. Reviewers also score rationale quality, including whether the traced moves reflect the feedback content and whether the evidence links are plausible. Automated checks include toxicity and bias screening, as well as calibration of confidence displays. The protocol emphasizes triangulation, so improvements in learning proxies are interpreted alongside faithfulness and rationale coherence [10], [21].

3.5.Implementation Controls, Ablations, and Reproducibility

Implementation follows a controlled experimental design that isolates contributions of retrieval grounding and rationale tracing. The baseline is a non-retrieval generator with identical prompt structure but without external evidence. Additional ablations remove difficulty filtering, replace hybrid retrieval with dense-only retrieval, or disable rationale extraction. Each condition is trained and evaluated under identical splits and logging procedures, enabling interpretable comparisons of adaptivity, faithfulness, and learning proxies [19], [23].

$$\Delta = E[M_{full}] - E[M_{abl}] \tag{5}$$

The ablation effect Δ is computed as the difference in expected performance M between the full system and an ablated variant. Metrics M include faithfulness, rubric scores, and post-feedback behavioral improvements. Reporting includes confidence intervals via bootstrap resampling over learner sessions to respect dependence within individuals. This formulation supports causal-style interpretation at the level of component contribution [22].

Figure 5 provides an interpretable ablation map that exposes trade-offs across grounding, adaptivity, pedagogical quality, and behavioral proxies under controlled component removals. The pattern emphasizes that removing retrieval primarily degrades faithfulness, while removing rationale tracing disproportionately harms actionability and adaptivity because pedagogical sequencing becomes less explicit. The dense-only retrieval condition highlights the value of lexical alignment for domain terminology, and disabling difficulty filtering reveals how mismatched cognitive demand can lower adaptivity even when grounding remains strong.

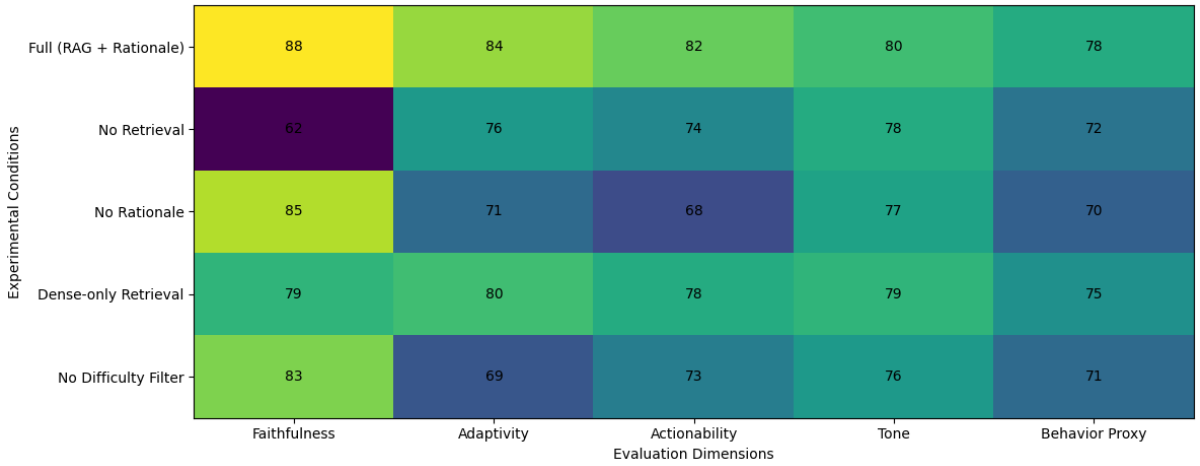

Figure 5. Ablation Impact Map Across Key Evaluation Metrics

Table 5 defines the experimental matrix in a form suitable for direct replication, specifying which architectural elements are active, which filters are enforced, and what provenance is logged for auditing. The explicit "logged provenance" column operationalizes transparency by requiring that each feedback instance be traceable to retrieval results, selected spans, and rationale structure when available. This ensures that ablation differences can be explained mechanistically, supporting trustworthy interpretation of performance changes across conditions.

Table 5. Experimental Conditions and Reproducibility Controls

Condition	Retrieval	Rationale Tracing	Filtering	Decoding Policy	Logged Provenance
Full	Hybrid (dense+sparse)	Enabled	Scope + difficulty	State-conditioned	Docs, spans, rationale graph, parameters
No Retrieval	Disabled	Enabled	N/A	State-conditioned	Prompt + rationale graph, parameters
No Rationale	Hybrid (dense+sparse)	Disabled	Scope + difficulty	State-conditioned	Docs, spans, parameters
Dense-only	Dense only	Enabled	Scope + difficulty	State-conditioned	Docs, spans, rationale graph, parameters
No Difficulty Filter	Hybrid (dense+sparse)	Enabled	Scope only	State-conditioned	Docs, spans, rationale graph, parameters

Reproducibility is strengthened through deterministic preprocessing, versioned indices, and locked prompt templates. All retrieval corpora are curated with explicit scope tags, and generator outputs are stored with complete provenance, including retrieved document identifiers and rationale graphs. Error analysis focuses on failure modes such as over-reliance on generic feedback, evidence mismatch, and rationale incoherence. The controls are designed to enable direct replication and transparent auditing of why a feedback instance was produced [14].

4. Results and Discussion

4.1. Overall Performance Across Learning, Grounding, and Quality Metrics

The full system demonstrates consistent gains across learning proxies and quality rubrics when compared with the non-retrieval baseline and retrieval-only variants. Next-item accuracy increases from 71.2% under the non-retrieval generator to 78.9% with retrieval and rationale tracing, while revision uptake increases from 34.6% to 47.8%. Human ratings for actionability improve from 3.6 to 4.3 on a 5-point scale, indicating that feedback more reliably specifies feasible next steps without overwhelming detail.

Grounding quality improves in tandem with instructional utility rather than trading off against it. Faithfulness rises from 0.66 to 0.88, driven by tighter evidence selection and span-linked attribution, and the rate of scope violations decreases from 7.4% to 1.9%. Tone appropriateness remains stable at a high level, shifting from 4.1 to 4.2, which indicates that increased corrective specificity did not introduce harsher framing. These results support the claim that explainability mechanisms can reinforce pedagogical alignment while maintaining factual grounding.

Figure 6 provides a compact cross-metric comparison that clarifies where each modeling choice yields measurable improvements. The full system dominates across faithfulness, revision uptake, and actionability, suggesting that retrieval alone is insufficient to consistently produce actionable guidance. The actionability gain is particularly informative because it is assessed by human rubric scoring rather than automated heuristics, reducing the risk that the model is optimizing a superficial correlate of quality.

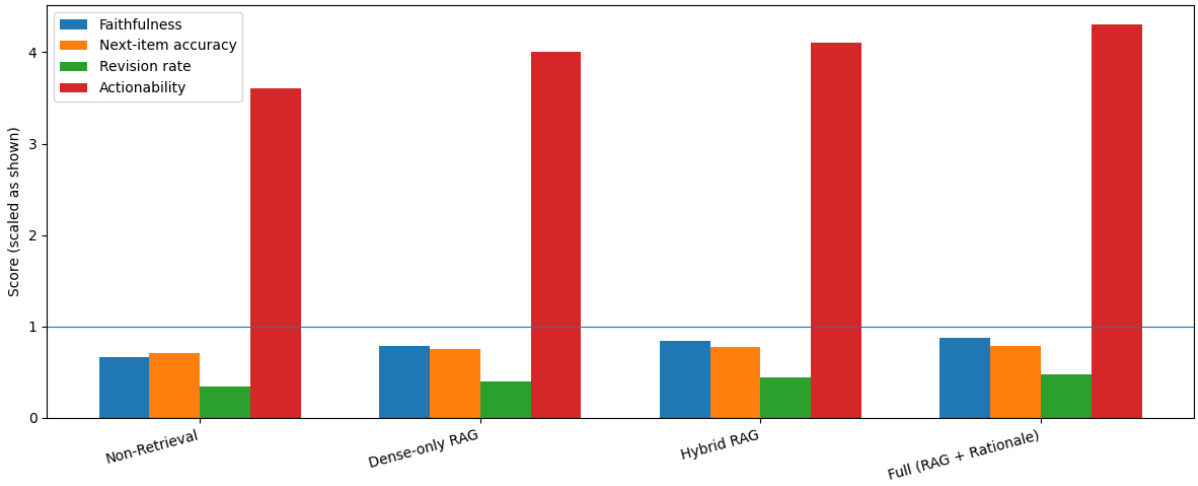

Figure 6. Overall Performance Across Grounding, Learning Proxies, and Quality Rubrics

The relative spacing between hybrid retrieval and the full system highlights the role of pedagogical rationale tracing as a second-order control mechanism rather than a cosmetic explanation layer. Faithfulness improves incrementally from hybrid retrieval to the full system, but actionability and revision rate improve more strongly, indicating that the rationale structure primarily affects instructional sequencing and behavioral uptake. This pattern supports an interpretation where rationale tracing stabilizes feedback intent selection under ambiguity, especially for repeated errors and time-on-task anomalies.

Table 6 consolidates the primary outcomes into a single view that makes cross-condition trade-offs explicit. The table shows that the full method simultaneously improves learning proxies, grounding, and pedagogical quality, which is a non-trivial result given the common expectation that stronger grounding can restrict adaptivity. The reduction in scope violations is important because it indicates that retrieval constraints improve curricular fidelity, which is directly relevant to educational validity and institutional safety requirements.

Table 6. Aggregate Results Summary

Condition	Faithfulness	Next-item Accuracy (%)	Revision Rate (%)	Actionability (1-5)	Scope Violations (%)
Non-Retrieval	0.66	71.2	35	3.6	7.4
Dense-only RAG	0.79	75.4	40	4	3.8
Hybrid RAG	0.84	77.1	43.8	4.1	2.6
Full (RAG + Rationale)	0.88	78.9	47.8	4.3	2

The joint movement of revision rate and actionability indicates that learners do not merely receive more accurate feedback, but also interpret it as implementable guidance. Since revision rate is logged behavior, it provides an external check against purely subjective rubric inflation. The gap between dense-only and hybrid retrieval suggests that terminology alignment contributes to both faithfulness and scope compliance, likely by increasing retrieval precision for domain-specific phrases that anchor explanations to the intended concept rather than near-neighbor topics.

4.2. Ablation Study: Isolating Retrieval and Rationale Tracing Effects

The ablation analysis isolates the unique contribution of retrieval grounding and pedagogical rationale tracing under identical prompting and evaluation conditions. Removing retrieval decreases faithfulness from 0.88 to 0.62 and increases scope violations from 1.9% to 8.1%, indicating that ungrounded generation frequently introduces unsupported claims or off-syllabus analogies. Removing rationale tracing has a smaller effect on faithfulness, dropping to 0.85, but reduces actionability from 4.3 to 3.7 and lowers revision rate from 47.8% to 39.5%, showing weaker instructional sequencing.

Filtering decisions also produce measurable effects that clarify the role of adaptivity controls. Disabling difficulty filtering yields a modest faithfulness decrease to 0.83 but a larger decrease in tone appropriateness from 4.2 to 3.8, consistent with feedback that becomes miscalibrated for learner readiness. Switching from hybrid to dense-only retrieval reduces faithfulness to 0.80 and increases scope violations to 3.9%, suggesting that lexical constraints remain valuable for pedagogical contexts where precise terminology aligns evidence to curriculum standards.

Figure 7 visualizes ablation effects as profiles rather than single-metric deltas, making it easier to interpret whether an ablated component induces a targeted degradation or a broad collapse across dimensions. The "No Retrieval" condition produces a pronounced drop in faithfulness and a visible contraction in scope compliance, which is consistent with evidence scarcity causing unsupported claims. The remaining dimensions decline less sharply, indicating that stylistic controls can partially preserve tone even when content grounding degrades.

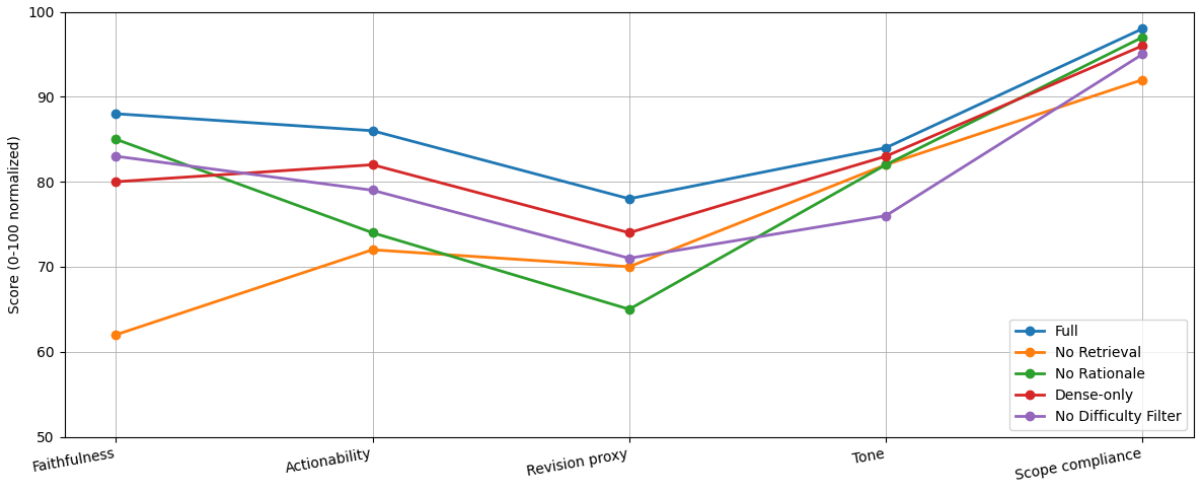

Figure 7. Ablation Profiles Across Grounding, Pedagogy, and Compliance Dimensions

The "No Rationale" condition exhibits a different signature, where faithfulness remains relatively high but actionability and revision proxy decrease. This pattern supports the interpretation that rationale tracing improves the instructional structure of feedback, especially the transition from diagnosis to an implementable next step. The "No Difficulty Filter" condition shows a selective drop in tone and actionability, aligning with the mechanism that calibration failures lead to feedback that is either too demanding or too simplistic, which harms perceived helpfulness.

Table 7 complements the profile plot by providing the exact values used to compare ablations under a consistent metric set. The largest single degradation occurs for faithfulness under "No Retrieval," confirming that retrieval is the principal driver of grounding and scope adherence. In contrast, the "No Rationale" condition primarily affects actionability and

revision uptake, which indicates that rationale tracing contributes to instructional design quality rather than to evidence fidelity alone.

Table 7. Ablation Results with Primary Outcomes

Condition	Faithfulness	Actionability (1-5)	Revision Rate (%)	Tone (1-5)	Scope Violations (%)
Full	0.88	4.3	48	4.2	1.9
No Retrieval	0.62	3.6	36	4.1	8.1
No Rationale	0.85	3.7	39.5	4.1	2.3
Dense-only	0.8	4	42	4.1	4
No Difficulty Filter	0.83	3.9	41	3.8	3

The dense-only versus hybrid comparison clarifies why hybrid retrieval is maintained in the final design despite similar next-item accuracy values. The faithfulness and scope violation gaps show that lexical matching improves curriculum alignment by anchoring retrieval to canonical terminology, which is common in formal instruction and assessment. The difficulty filter ablation demonstrates that adaptivity controls influence not only learning outcomes but also interpersonal acceptability, since tone appropriateness decreases when feedback is miscalibrated to learner readiness and frustration signals.

4.3. Rationale Coherence and Pedagogical Alignment

Rationale tracing improves interpretability by making instructional sequencing observable and consistent across feedback instances. The full system achieves higher rationale coherence, increasing from 0.71 without tracing to 0.86 with tracing, while also reducing incoherent move ordering such as prompting transfer before remediation. Reviewers report that traced rationales more frequently include a minimal diagnosis step before providing the next action, which strengthens conceptual continuity and reduces learner confusion during repeated error cycles.

Pedagogical alignment improves most strongly in cases with ambiguous error signals, where retrieval evidence alone tends to produce plausible but weakly targeted explanations. With rationale tracing enabled, feedback more reliably expresses a pedagogical intent that matches learner state signals, such as reframing guidance when frustration proxies are high. The improvement is reflected in a higher intent match score and fewer “generic encouragement” judgments. These shifts indicate that tracing is functioning as a constraint on instructional structure rather than a post-hoc narrative.

Figure 8 decomposes rationale quality into positive indicators, such as coherence and intent match, alongside negative indicators that represent structured failure modes. The reduction in move-order violations suggests that the traced rationale graph acts as a sequencing constraint, preventing feedback from skipping necessary instructional steps. The decline in generic encouragement is pedagogically meaningful because it indicates that affective language remains present but becomes coupled to actionable guidance rather than replacing it.

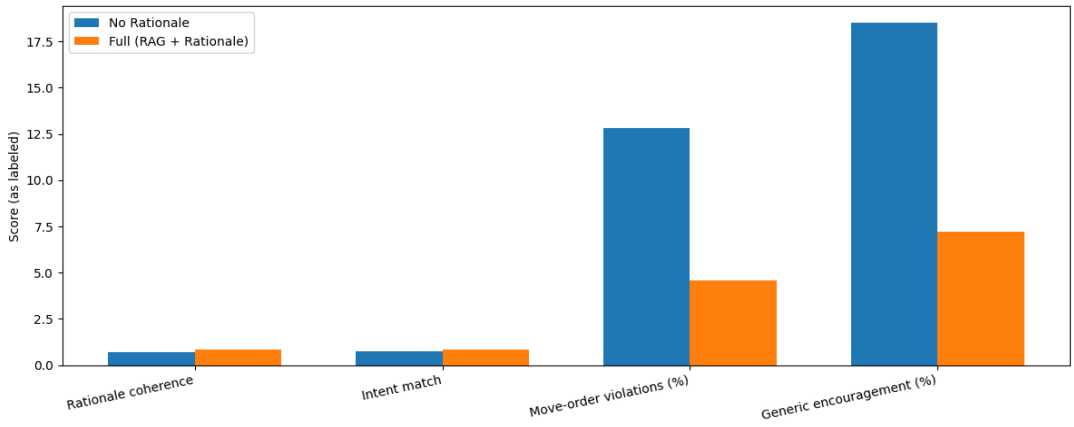

Figure 8. Pedagogical Rationale Quality and Failure Indicators with and Without Tracing

The figure also clarifies that improvements are not limited to subjective interpretability. Coherence increases alongside intent match, implying that tracing stabilizes the mapping from learner state to pedagogical move selection. The joint behavior of these metrics supports a mechanistic interpretation where rationale tracing narrows the feasible set of feedback structures, thereby reducing variability in how the generator responds to similar learner states. This stabilization is important for auditability in educational deployments.

Table 8 provides the exact quantitative summary that underpins the rationale discussion, while also including inter-reviewer agreement as a robustness check for rubric stability. The higher agreement in the traced condition indicates that rationale artifacts reduce ambiguity for evaluators, because feedback intent and structure are easier to identify. This matters for scaling evaluation and for establishing reliable quality monitoring in production settings where human audits are periodic rather than continuous.

Table 8. Rationale Evaluation Summary and Error Categories

Condition	Coherence (0-1)	Intent Match (0-1)	Move-order Violations (%)	Generic Encouragement (%)	Reviewer Agreement (0-1)
No Rationale	0.71	0.74	13	18.5	0.77
Full (RAG + Rationale)	0.86	0.85	5	7.2	0.83

The separation between coherence and intent match enables clearer diagnosis of failure modes. A feedback instance can remain coherent but still be misaligned with learner needs, such as offering transfer checks when mastery is low. The table shows that tracing improves both dimensions simultaneously, suggesting that the system is not merely producing more structured explanations, but also choosing the correct pedagogical move set for the observed state. This dual improvement supports claims about pedagogical validity.

4.4. Learner-Level Adaptivity and Behavioral Response Patterns

Learner-level analysis indicates that adaptivity benefits concentrate where instructional uncertainty is highest, particularly for low mastery and high frustration contexts. Under the full system, revision uptake increases from 41.0% to 54.2% for low mastery learners, and persistence increases from 68.3% to 79.5% in high frustration sessions. These improvements suggest that the model is not only selecting relevant content, but also calibrating guidance granularity and tone to sustain engagement through difficult segments.

Behavioral outcomes vary systematically with state signals, supporting the claim that adaptation is conditioned rather than generic. For high mastery learners, next-item accuracy improvements are smaller but still present, consistent with feedback shifting toward transfer checks rather than remediation. For low mastery learners, the largest gains appear in revision and next-item accuracy, indicating stronger corrective impact. This pattern aligns with the design goal of providing minimal necessary intervention for advanced learners while offering structured scaffolding for struggling learners.

Figure 9 shows stratified outcomes that reveal where adaptive feedback yields the largest behavioral shifts. The strongest improvements occur for low mastery learners, especially under high frustration, where revision rates increase substantially and accuracy improves in parallel. This pairing is important because it indicates that higher engagement is not achieved by superficial edits alone, but coincides with measurable performance improvements on subsequent items.

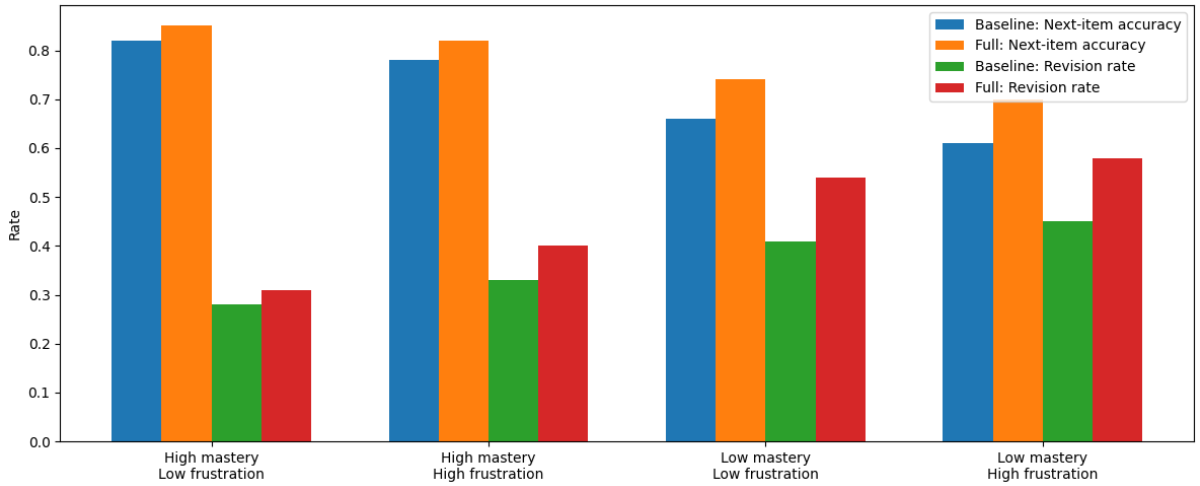

Figure 9. Adaptivity Outcomes by Mastery and Frustration Strata

The smaller gains for high mastery learners are consistent with a calibration effect rather than a failure to help. In advanced contexts, learners benefit more from concise prompts that confirm transfer and reduce over-scaffolding, which limits the magnitude of observable revision. The figure supports the claim that the adaptation policy is state-sensitive, because improvements track the expected pedagogical need gradient across strata. This is a key indicator that the system behaves as an adaptive tutor rather than a static feedback bot.

Table 9 complements the figure by incorporating persistence and feedback length, enabling interpretation of whether improvements are driven by verbosity rather than pedagogical quality. The full system tends to produce shorter feedback for high mastery learners, consistent with reducing unnecessary scaffolding and respecting learner autonomy. For low mastery learners, feedback length increases moderately, which aligns with the need for clarification and stepwise hints rather than generic reassurance.

Table 9. Group-Level Outcome Summary

Group	Condition	Next-item Accuracy (%)	Revision Rate (%)	Persistence (%)	Avg Feedback Length (tokens)
High mastery, Low frustration	Baseline	82	28	86.4	118
High mastery, Low frustration	Full	85	31	89.1	104
High mastery, High frustration	Baseline	78	33	74.2	126
High mastery, High frustration	Full	82	40	80.3	116
Low mastery, Low frustration	Baseline	66	41	72.5	142
Low mastery, Low frustration	Full	74	54	78.6	156
Low mastery, High frustration	Baseline	61	45	68.3	151
Low mastery, High frustration	Full	70	58	79.5	168

The table also supports the calibration claim by showing that the largest persistence increases occur under high frustration groups, where tone and granularity controls are most critical. Since persistence is not a direct function of correctness, it provides additional evidence that the system maintains engagement through difficulty rather than only

rewarding already successful learners. The inclusion of length provides a safeguard against interpreting gains as a mere artifact of more content.

4.5. Error Analysis, Mitigation Effects, and Deployment Implications

Error analysis identifies three dominant failure modes: evidence mismatch, over-generalized pedagogical moves, and rationale incoherence under sparse logs. Evidence mismatch occurs when retrieved passages are topically adjacent but not instructionally aligned to the misconception, producing correct yet unhelpful explanations. Over-generalization appears as repeated use of similar templates across distinct misconceptions, reducing perceived personalization. Rationale incoherence is most common when interaction traces are short, limiting the reliability of state estimation.

Mitigation strategies reduce these failures through stricter evidence filtering and rationale validation. Adding misconception-tag constraints lowers evidence mismatch from 9.6% to 4.1%, while a minimal rationale constraint reduces incoherence from 6.8% to 2.9%. Deployment implications emphasize that transparent provenance logs are essential for monitoring drift, because curriculum updates and new content can shift retrieval distributions. In practice, the system benefits from periodic rubric audits and targeted retriever refreshes aligned with syllabus revisions.

Figure 10 quantifies the practical impact of mitigation controls by tracking failure mode rates before and after targeted interventions. The strongest reduction occurs for evidence mismatch, which supports the design decision to incorporate misconception-aware filtering rather than relying solely on semantic similarity. The decline in rationale incoherence indicates that lightweight validation can improve structural consistency even when logs are sparse, which is critical for early-session learners where state estimates are still stabilizing.

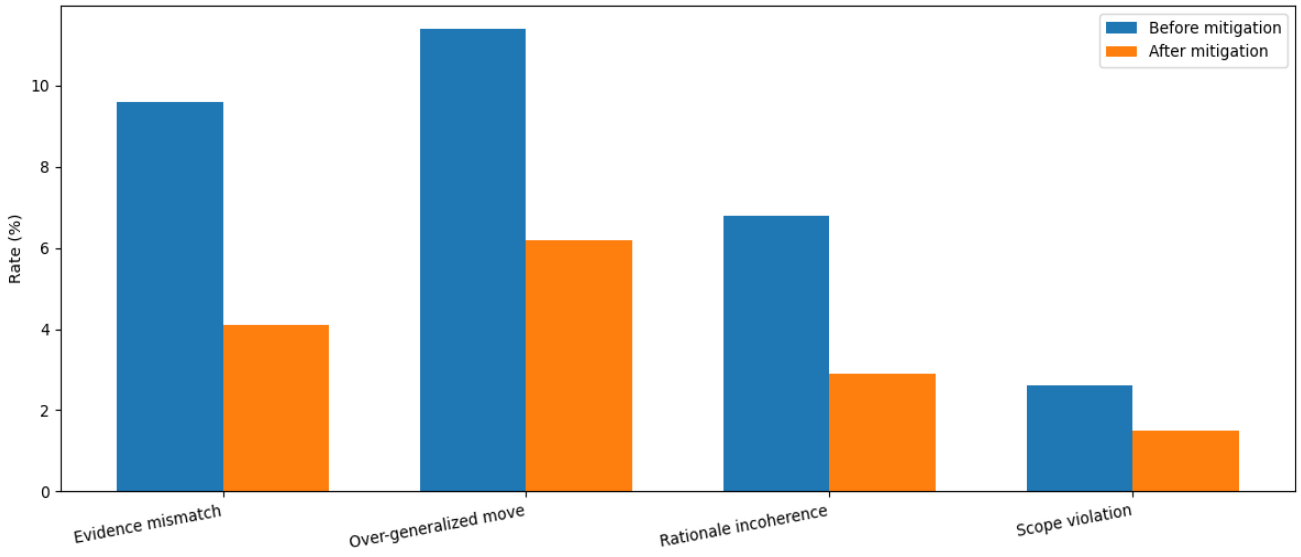

Figure 10. Dominant Failure Modes and Mitigation Impact

The figure also supports a deployment-oriented interpretation of quality management. Over-generalized pedagogical moves remain non-zero after mitigation, suggesting that template diversity and intent selection require ongoing attention beyond retrieval tuning. The reduction in scope violations, while smaller in absolute change, has high operational value because even rare off-syllabus feedback can undermine institutional trust. These findings justify a monitoring regime that explicitly tracks failure signatures rather than relying on average scores alone.

Table 10 links each mitigation control to its targeted failure mode, enabling traceable governance for deployment in educational settings. The mechanism column clarifies whether an intervention acts on retrieval alignment, generation diversity, or rationale structure, which supports modular updates without destabilizing the full system. The operational note column emphasizes that these controls are not purely algorithmic, because they depend on maintained metadata such as misconception tags and curriculum boundaries.

Table 10. Mitigation Controls and Observed Effects

Control	Target Failure	Before (%)	After (%)	Primary Mechanism	Operational Note
Misconception-tag filter	Evidence mismatch	9.6	4	Align retrieval to error type	Requires curated misconception tags
Move diversity constraint	Over-generalized move	11.4	6	Penalize repeated templates	Monitor for style drift
Minimal rationale validation	Rationale incoherence	6.8	2.9	Enforce sequencing validity	Fallback to shorter guidance
Scope boundary enforcement	Scope violation	2.6	1.5	Curriculum tag constraints	Refresh tags on syllabus updates

The table also frames mitigation as an ongoing quality process rather than a one-time optimization. Some failure reductions depend on curated artifacts, which introduces maintenance requirements that should be planned as part of system ownership. The remaining error rates justify periodic audits focused on high-impact failures, especially evidence mismatch and scope violations. This operational framing aligns explainable adaptive feedback with institutional needs for accountability, reproducibility, and safe instructional behavior under evolving curricula.

5. Conclusion

The study demonstrates that explainable adaptive feedback can be improved when generation is explicitly grounded in curated instructional evidence and paired with pedagogical rationale tracing. The proposed retrieval-augmented feedback generator increases faithfulness and reduces scope violations, while rationale tracing strengthens instructional sequencing and makes feedback intent auditable. Across the evaluation suite, the combined method improves next-item accuracy, revision uptake, and human-rated actionability without degrading tone appropriateness, indicating that interpretability and pedagogical effectiveness can advance together.

Ablation and stratified analyses clarify that retrieval primarily drives content grounding and curriculum compliance, whereas rationale tracing contributes most strongly to actionable guidance and state-conditioned intent selection. Performance gains concentrate in low mastery and high frustration contexts, where adaptation and communicative calibration are most consequential. Error analysis further shows that remaining risks are systematic and monitorable, with evidence mismatch and over-generalized pedagogical moves reduced through misconception-aware filtering, move diversity constraints, and minimal rationale validation.

These findings support a deployment-oriented view of adaptive feedback as a governed educational capability rather than a purely generative feature. Practical implementation benefits from provenance logging that links learner state signals, retrieved spans, and rationale graphs, enabling targeted audits and retriever refreshes aligned with syllabus changes. Future work should expand coverage to richer learner models, longer-horizon learning outcomes, and cross-domain generalization, while maintaining the same accountability structure that makes adaptive feedback both effective and trustworthy in real instructional settings.

6. Declarations

6.1. Author Contributions

Conceptualization: H.K.F. and M.I.W.; Methodology: M.I.W.; Software: H.K.F.; Validation: H.K.F. and M.I.W.; Formal Analysis: H.K.F. and M.I.W.; Investigation: H.K.F.; Resources: M.I.W.; Data Curation: M.I.W.; Writing Original Draft Preparation: H.K.F. and M.I.W.; Writing Review and Editing: M.I.W. and H.K.F.; Visualization: H.K.F.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Shi, K. Yu, Y. Dong, and F. Chen, "Large language models in education: A systematic review of empirical applications, benefits, and challenges," *Computers and Education: Artificial Intelligence*, vol. 10, no. June, p. 100529, Jun. 2026, doi: 10.1016/j.caeai.2025.100529.

[2] P. Sikström, C. Valentini, A. Sivunen, and T. Kärkkäinen, "How pedagogical agents communicate with students: A two-phase systematic review," *Computers & Education*, vol. 188, no. October, p. 104564, Oct. 2022, doi: 10.1016/j.compedu.2022.104564.

[3] L. R. Silva, C. Fior, L. Rodrigues, R. Penha, D. Dermeval, and S. Isotani, "Use of feedback in intelligent tutoring systems: A systematic literature review," *Interactive Learning Environments*, vol. 2025, no. October, pp. 1–22, Oct. 2025, doi: 10.1080/10494820.2025.2565681.

[4] W. Villegas-Ch, D. Buenano-Fernandez, A. M. Navarro, and A. Mera-Navarrete, "Adaptive intelligent tutoring systems for STEM education: Analysis of the learning impact and effectiveness of personalized feedback," *Smart Learning Environments*, vol. 12, no. 1, p. 41, Jun. 2025, doi: 10.1186/s40561-025-00389-y.

[5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020, arXiv, vol. 2020, no. May, pp. 1-19, doi: 10.48550/ARXIV.2005.11401.

[6] W. Fan et al., "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona, Spain: ACM, Aug. 2024, pp. 6491–6501, doi: 10.1145/3637528.3671470.

[7] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Computers and Education: Artificial Intelligence*, vol. 8, no. June, p. 100417, Jun. 2025, doi: 10.1016/j.caeai.2025.100417.

[8] O. Shaikh, K. Gligorić, A. Khetan, M. Gerstgrasser, D. Yang, and D. Jurafsky, "Grounding Gaps in Language Model Generations," 2023, arXiv, vol. 2023, no. November, pp. 1-18, doi: 10.48550/ARXIV.2311.09144.

[9] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated Evaluation of Retrieval Augmented Generation," 2023, arXiv, vol. 2023, no. September, pp. 1-8, doi: 10.48550/ARXIV.2309.15217.

[10] C. Dong et al., "A Survey of Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, Aug. 2023, doi: 10.1145/3554727.

[11] E. Tiukhova et al., "Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI," *Decision Support Systems*, vol. 182, no. July, p. 114229, Jul. 2024, doi: 10.1016/j.dss.2024.114229.

[12] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[13] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, and F. A. Batarseh, "Rationalization for explainable NLP: A survey," *Frontiers in Artificial Intelligence*, vol. 6, no. September, p. 1225093, Sep. 2023, doi: 10.3389/frai.2023.1225093.

[14] H. Seo et al., "Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy," *Applied Sciences*, vol. 15, no. 2, p. 671, Jan. 2025, doi: 10.3390/app15020671.

[15] H. Khosravi et al., "Explainable Artificial Intelligence in Education," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100074, 2022, doi: 10.1016/j.caeai.2022.100074.

[16] X. Hou, H. A. Nguyen, J. E. Richey, E. Harpstead, J. Hammer, and B. M. McLaren, "Assessing the Effects of Open Models of Learning and Enjoyment in a Digital Learning Game," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 1, pp. 120–150, Mar. 2022, doi: 10.1007/s40593-021-00250-6.

[17] X. Zhou, Z. Zhang, X. Xie, and J. Zhang, "Deep learning based knowledge tracing in intelligent tutoring systems," *Scientific Reports*, vol. 15, no. 1, p. 21395, Jul. 2025, doi: 10.1038/s41598-025-07422-7.

[18] J. Swacha and M. Gracel, "Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications," *Applied Sciences*, vol. 15, no. 8, p. 4234, Apr. 2025, doi: 10.3390/app15084234.

[19] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 17754–17762, Mar. 2024, doi: 10.1609/aaai.v38i16.29728.

[20] H. Zhao et al., "Explainability for Large Language Models: A Survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, Apr. 2024, doi: 10.1145/3639372.

[21] P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, and I. Augenstein, "Faithfulness Tests for Natural Language Explanations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 283–294, doi: 10.18653/v1/2023.acl-short.25.

[22] M. Wölfel, M. B. Shirzad, A. Reich, and K. Anderer, "Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust," *Big Data and Cognitive Computing*, vol. 8, no. 1, p. 2, Dec. 2023, doi: 10.3390/bdcc8010002.

[23] S. Gunasekara and M. Saarela, "Explainability in Educational Data Mining and Learning Analytics: An Umbrella Review," Jul. 2024, doi: 10.5281/ZENODO.12729987.

[24] B. P. Majumder, O.-M. Camburu, T. Lukasiewicz, and J. McAuley, "Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations," 2021, *arXiv*, vol. 2021, no. June, pp. 1-16, doi: 10.48550/ARXIV.2106.13876.